Using Generating Functions to Prove Additivity of Gene-Neighborhood Based Phylogenetics - Extended Abstract

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

[1] Department of Mathematics, Ort Braude, Israel
[2] Department of Evolutionary and Environmental Biology, University of Haifa, Israel
[3] RICAM, Austrian Academy of Sciences, Linz, Austria
[4] Department of Mathematics, Rutgers University, USA
[5] School of Mathematics and Statistics, University of Canterbury, NZ
[6] Department of Evolutionary and Environmental Biology, University of Haifa, Israel

Abstract. Prokaryotic evolution is often described as the *Spaghetti of Life* due to massive genome dynamics (GD) events of gene gain and loss, resulting in different evolutionary histories for the set of genes comprising the organism. These different histories, dubbed as *gene trees* provide confounding signals, hampering the attempt to reconstruct the *species tree* describing the main trend of evolution of the species under study.
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. Recently, GD has been modelled as a continuous-time Markov process. Under this formulation, the distance between genes along the chromosome was shown to follow a birth-death-immigration process. Using classical results from birth-death theory, we recently showed that the SI measure is consistent under that formulation.
In this work, we provide an alternative, stand alone combinatorial proof of the same result. By using generating function techniques we 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 the expected distances between organisms based on a transformation of their SI. Although the expressions obtained are rather complex, we establish additivity of this estimated evolutionary distance (a desirable property yielding phylogenetic consistency). This approach relies on holonomic functions and the Zeilberger Algorithm in order to establish additivity of the transformation of SI.

Keywords: Genome Dynamics, Markovian Processes, Generating Functions, Phylogenetics, Holonomic Functions.

1 Introduction

The dramatic advancements in sequencing technologies have made realistic biological tasks seemed imaginary only a decade ago. Inferring the evolutionary

history of thousands of species, is among the most fundamental tasks in biology with implications to medicine, agriculture, and more. Such a history is depicted in a tree structure and is called a *phylogeny*. Leaves of that tree correspond to contemporary (i.e. extant) species and the tree edges (or branches) represent evolutionary relationships. Despite the impressive advancement in the extraction of such molecular data, and of ever increasing quality, finding the underlying phylogenetic 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. The tree sought should preserve the property that the length of the path between any two organisms at its leaves equals the inferred pairwise distance between these two organisms. When such a tree exists, these distances are called *additive*, as does the distance matrix storing them.
Modern approaches in systematics rely on statistical modelling in which a model fitting optimally the data is sought. The challenges under this framework, are both statistical, i.e. accurately modelling the data, and computational for efficient model inference and selection from given data. In phylogenetics, *maximum likelihood* seeks for a tree under which the probability of observing the given leaf sequences is maximised [12,13,14,8,9]. Normally, the data for this task is taken from few ubiquitous genes, such as ribosomal genes, that reside in every species and are imune for GD events. Such genes are typically highly conserved by definition and hence cannot provide a strong enough signal to distinguish the shallow branches of the prokaryotic tree. Nevertheless, GD events, gene gain in the form of horizontal gene transfer (HGT), a mechanism by which organisms transfer genetic material not through vertical inheritance, and gene loss, seem to provide valuable evolutionary information that can be harnessed for classification [7,20,23]. Approaches relying on GD are mainly divided into gene-order-based and gene-content-based techniques. Under the gene-order-based approach [24,11,33], 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 gene-content-based approach [29,30,10] ignores entirely gene order, and similarity is defined as the size of the set of shared genes. Although a statistical framework was devised for part of these models [26,31,4,25] to the best of our knowledge no such framework accounted for HGT.

The *synteny index* (SI) [28,1,27] captures both existence and locality, i.e. gene content and order respectively, by summarising gene neighbourhoods across the entire genome. An attractive property of the SI measure is the relaxation of the equal gene content requirement, in which genomes are permutations of the gene set. Under the attempt to model SI in a statistical framework, the *Jump model* was defined to account for gene order variation between evolving genomes. The *Jump operation* moves a gene to a random location in the genome. In the *Jump model*, every gene jumps, in a Poisson process. Under that framework, a genome is defined as a continuous-time Markov process (CTMP) [2]. Consequently, gene distance along the genome 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. Therefore, trees were constructed from evolutionary distances inferred heuristically based on exponential decay modelling.

In a recent paper [19] we have used classical tools for the birth-death field such as spectral theory and orthogonal polynomials, to derive analytical expressions for deriving the model transition probabilities and hence expected evolutionary distances. These analytical expressions yielded *model consistency* - an attractive property in systematics, implying that a measure infers accurate distances under a given model of evolution.

In this work, we provide an alternative, standalone combinatorial derivation for the model parameter and the proof of consistency. We first define the system in terms of a generating function, and extract transition probabilities as a function of time since divergence. However, the complexity of the expressions obtained to infer distances, could not readily imply consistency for the SI. By showing that these expressions satisfy the conditions for holonomic functions [35,32] and applying the Zeilberger Algorithm [34] we prove consistency of the SI measure under the jump model. We believe that this alternative proof, besides its independent interest, confers better understandings of the system and might prove useful for future extensions of the model, handling richer models such as unequal gene content or jumps of several genes.
Due to space considerations, several proofs were omitted and will appear at the journal version.

2 Preliminaries

We provide preliminary definitions and concepts to be used throughout the paper. We start with the Jump Model that comprises of a Jump operation operating on a single gene, and a stochastic process acting on the genomic ensemble of genes.

The *Jump* Model In this work we consider the genome as a gene list, that is, the basic unit of resolution is an entire gene. Let $\mathcal{G}^{(n)} = (g_1, g_2, \ldots, g_n)$ be a sequence of 'genes' (see Figure 1). For the sake of ignoring the tips of the sequence $\mathcal{G}^{(n)}$, we assume n is large enough compared to other sizes defined below.

Let $\mathcal{G}^{(n)}(0)$ be a genome at time zero and WLOG let $\mathcal{G}^{(n)}(0) = (g_1, g_2, \ldots, g_n)$. Now 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 (or simply *Jumps*) to a different location in the sequence, with each of these possible $n-1$ locations selected uniformly at random (see Fig 2).

For example, if $\mathcal{G}^{(n)}(t) = (g_1, g_2, g_3, g_4, g_5)$, and then g_1 jumps and lands between g_3 and g_4 then the sequence yielded is $\mathcal{G}^{(n)}(t+\delta) = (g_2, g_3, g_1, g_4, g_5)$. Note, that g_i can also move to one of the tips of the genome.

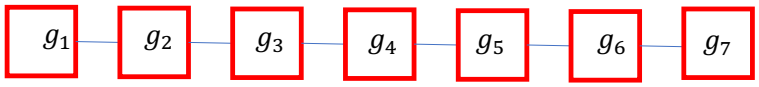

Fig. 1. A Genome as Gene List: The basic unit of resolution is a gene and a genome is defined as a sequence of genes.

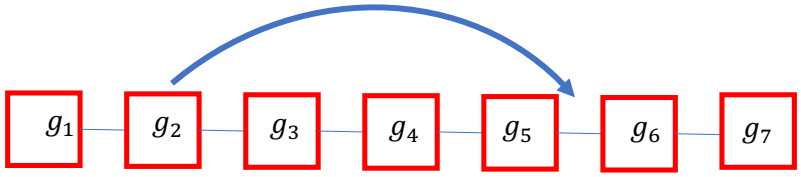

Fig. 2. The Jump operation: Gene g_2 jumps into the space between genes g_5 and g_6.

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 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.

The Synteny Index Let k be any constant positive integer (note it may be possible to allow k to grow slowly with n but we will not explore such an extension here). Then, for $j \in k+1, \ldots, n-k$ the $2k$-neighbourhood of gene g_j in a genome $\mathcal{G}^{(n)}$, $N_{2k}(g_j, \mathcal{G}^{(n)})$ is the set of $2k$ genes (different from g_j) that have distance, in terms of separating genes along the chromosome, at most k from g_j in $\mathcal{G}^{(n)}$. Consider genomes $\mathcal{G}_1^{(n)}$ and $\mathcal{G}_2^{(n)}$, with the restriction that $\mathcal{G}_1^{(n)}$ and $\mathcal{G}_2^{(n)}$ share the same gene set. Let $SI_j(\mathcal{G}_1^{(n)}, \mathcal{G}_2^{(n)})$ be the relative intersection size between $N_{2k}(g_j, \mathcal{G}_1^{(n)}))$ and $N_{2k}(g_j, \mathcal{G}_2^{(n)})$, or formally

$$SI_j(\mathcal{G}_1^{(n)}, \mathcal{G}_2^{(n)}) = \frac{1}{2k}|N_{2k}(g_j, \mathcal{G}_1^{(n)}) \cap N_{2k}(g_j, \mathcal{G}_2^{(n)})|$$

(this is also called *the Jaccard index* between the two neighbourhoods [15]). See Figure 3 for example of a gene neighbourhood and the synteny index of a particular gene.

For the special case of our stochastic process, we define $SI_j(t)$ to be the SI for gene g_j between $\mathcal{G}^{(n)}(0)$ and $\mathcal{G}^{(n)}(t)$, $\frac{1}{2k}|N_{2k}(g_j, \mathcal{G}^{(n)}(0)) \cap N_{2k}(g_j, \mathcal{G}^{(n)}(t))$.

Let $\overline{SI}(\mathcal{G}_1^{(n)}, \mathcal{G}_2^{(n)})$ be the average of these $SI_j(\mathcal{G}_1^{(n)}, \mathcal{G}_2^{(n)})$ values over all genes g_j for j between $k+1$ and $n-k$. That is,

$$\overline{SI}(\mathcal{G}_1^{(n)}, \mathcal{G}_2^{(n)}) = \frac{1}{n-2k} \sum_{j=k+1}^{n-k} SI_j(\mathcal{G}_1^{(n)}, \mathcal{G}_2^{(n)}).$$

Finally, we equivalently define $\overline{SI}(\mathcal{G}^{(n)}(0), \mathcal{G}^{(n)}(t)$ be the average of these $SI_j(t)$ values between $\mathcal{G}^{(n)}(0)$ and $\mathcal{G}^{(n)}(t)$, over all j from $k+1$ to $n-k$.

$$\overline{SI}(\mathcal{G}_1^{(n)}, \mathcal{G}_2^{(n)}) = \frac{1}{n-2k} \sum_{j=k+1}^{n-k} SI_j(t). \tag{1}$$

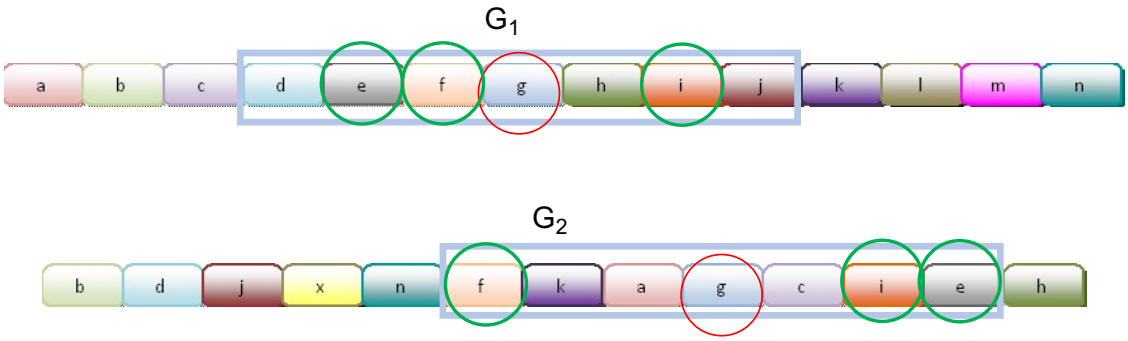

Fig. 3. The synteny Index The two gene neighbourhoods induced by gene g in genomes G_1 and G_2 and the synteny Index between G_1 and G_2 for gene g, $SI_g(\mathcal{G}_1^{(n)}, \mathcal{G}_2^{(n)}) = \frac{1}{2k}|N_{2k}(g, \mathcal{G}_1^{(n)}) \cap N_{2k}(g, \mathcal{G}_2^{(n)})|$. As genes e, f and i are shared between the two neighbourhoods induced by gene g in G_1 and G_2, we obtain $SI_g(\mathcal{G}_1^{(n)}, \mathcal{G}_2^{(n)}) = \frac{1}{2}$.

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

2.1 Genome Permutations as a State Space

We now introduce a random process, that will play a key role in the analysis of the random variable $\overline{SI}(\mathcal{G}^{(n)}(0), \mathcal{G}^{(n)}(t))$. Consider the location of a gene g_i, not being transferred during time period t, with respect to another gene $g_{i'}$. WLG

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 $g_{i'}$ one position away from g_i, and transfer from that interval, moves $g_{i'}$ 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 (generalised linear) birth-death process, and the process is illustrated in Fig 4. As the process is not affected by the specific values of i and i' (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.

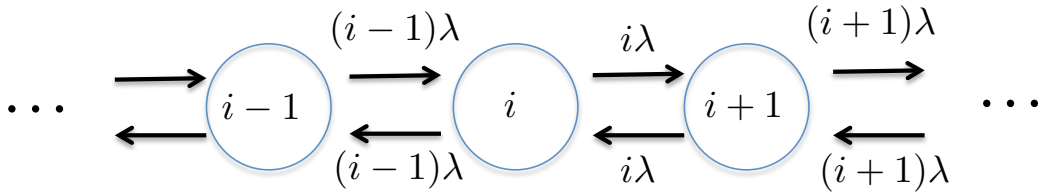

Fig. 4. The Markov Chain as a Birth-Death process Transitions between the states in the linear birth-death process with linear rate's growth/decrease.

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 divergence times that we establish below.

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. Along this work we will use the term length as it corresponds to number of events or time span. Edge lengths are naturally extended to *paths* where path length is the sum of edge lengths along the path. For a tree T over n leaves, let $D(T)$ (or simply D) be a symmetric $n \times n$ matrix where $[D]_{i,j}$ holds the path length (distance) between leaves i and j in T. A matrix D' is called *additive* if there is a tree T' such that $D(T') = D'$. A distance measure is considered *additive on a model* M if it can be transformed (or *corrected*) to the expected number of events generated under M.

3 Asymptotic Estimation of the Model Parameters

In order to reconstruct maximum likelihood trees, we need to estimate the model parameters, in a way that maximises their likelihood. We here establish the main theoretical result of this work, by defining the problem parameters as a generating function and use rules from this area. That in turn yields an analytical expression of divergence times. Recall that we wish to link SI to our model parameter X_t which is the expected value (state) of the model. Such a linkage was established in [27] that we restate explicitly below. While this expression is essential for the analysis, it is not stated in terms of the parameters of the model, specifically the time since divergence, and therefore has limited power.
We start with some essential definitions that are central in the analysis. 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. Formally,

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

$p_{i,j}(t)$ is the most basic variable and on which more special variables are defined. Now denote

$$q_{i,k}(t) = \sum_{j=1}^{k} p_{i,j}(t) \tag{2}$$

the conditional probability that $X_t \leq k$ given that $X_0 = i$, as $q_{i,k}(t)$.

Also let $q_k(t)$ denote 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, or formally:

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

Having defined these variables, we can restate the fundamental theorem we proved in [27]:

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

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

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

Theorem 1 is important as it links between SI, event rate, and probabilities of genes staying at their original neighbourhoods. Nevertheless, these factors are confounded in the sense that $q_k(t)$ depends on t, and therefore it would be desirable to arrive at an expression stated in the parameters of the model, i.e. time and rate solely, so divergence times, or alternatively number of events, can be estimated and trees can be reconstructed. The rest of the section is devoted to this.

3.1 Finding the Model Transition Probabilities

The transition probabilities of the Markov Model defined above are fundamental for our goal - analytical expressions of the expected synteny index between two genomes in terms of their divergence time. Hence, finding an explicit expression, in terms of i, j, and t, is our first task.

Theorem 2.

$$p_{i,j}(t) = \frac{1}{(t+1)^{i+j-1}} \cdot \sum_{\ell=1}^{\min(i,j)} \frac{(i+j-\ell-1)!}{(i-\ell)!(j-\ell)!(\ell-1)!} \left(1-t^2\right)^{\ell-1} t^{i+j-2\ell}. \quad (4)$$

The next step uses a lemma from [27] that adapts the Forward Kolmogorov Equation [2] to our special setting.

Lemma 1. *[27]*

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

$$\frac{1}{\lambda}\frac{dp_{i,j}(t)}{dt} = -(2j-1)p_{i,j}(t) + jp_{i,j+1}(t) + (j-1)p_{i,j-1}(t), \quad (5)$$

subject to the initial condition:

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

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

$$\mathbb{E}[X_t \mid X_0 = i] = i + t\lambda. \quad (6)$$

Our first aim is to solve the above infinite system of differential equations (5).

Without loss of generality, we assume $\lambda = 1$ and introduce the following definition making use of a generating function.

Definition 2. *For each $i \geq 1$, we define a generating function $f_i(t,z) = \sum_{j=1}^{\infty} p_{i,j}(t) z^j$.*

Lemma 2.

$$f_i(t,z) = \frac{t^{i-1}}{(t+1)^i} \sum_{j=1}^{\infty} (-1)^{j-1} z^j \left(\frac{t}{t+1}\right)^{j-1} \sum_{\ell=1}^{\min(i,j)} \binom{i-1}{\ell-1}\binom{-i}{j-\ell}\left(\frac{t^2-1}{t^2}\right)^{\ell-1} \quad (7)$$

The full proof of Lemma 2 is deferred to the journal version.

We are now in a position to prove Theorem 2.

Proof of Theorem 2: From Definition 2 and Lemma 2 above, we have two equal power series. Therefore, by the uniqueness of generating functions [22], we conclude that the coefficients are pairwise equal, hence:

$$\begin{aligned} p_{i,j}(t) &= (-1)^{j-1}\frac{t^{i-1}}{(t+1)^i}\left(\frac{t}{t+1}\right)^{j-1}\cdot\sum_{\ell=1}^{\min(i,j)}\binom{i-1}{\ell-1}\binom{-i}{j-\ell}\left(\frac{t^2-1}{t^2}\right)^{\ell-1} \\ &= (-1)^{j-1}\left(\frac{1}{t+1}\right)^{i+j-1}\cdot\sum_{\ell=1}^{\min(i,j)}\binom{i-1}{\ell-1}\binom{-i}{j-\ell}\left(t^2-1\right)^{\ell-1}t^{i+j-2\ell} \\ &= \left(\frac{1}{t+1}\right)^{i+j-1}\cdot\sum_{\ell=1}^{\min(i,j)}\binom{-i}{j-\ell}\frac{(i-1)!}{(i-\ell)!(\ell-1)!}(-1)^{j-1}\left(t^2-1\right)^{\ell-1}t^{i+j-2\ell} \end{aligned} \tag{8}$$

Recalling that the generalisation of the binomial coefficient to negative integers $-n$ is:

$$\binom{-n}{k} = (-1)^k\binom{n+k-1}{k}$$

we obtain from (8):

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

□

From Theorem 2 we can see the following:

Corollary 1. *For any i, j, and t it holds that $p_{i,j}(t) = p_{j,i}(t)$.*

3.2 Expectation and Variance of X_t

Having explicit expression for $p_{i,j}(t)$ allows us to confirm other derived values. Therefore we here note by passing the expected value and variance of X_t. By the definition of X_t we have

$$E(X_t \mid X_0 = i) = \sum_{j=1}^{\infty} j p_{i,j}(t)$$

.

Also, from Definition 2 we have

$$\frac{d}{dz}f_i(t,z) = \frac{d}{dz}\sum_{j=1}^{\infty}p_{i,j}(t)z^j = \sum_{j=1}^{\infty}jp_{i,j}(t)z^{j-1}$$

Using the generating functions we have

$$E(X_t \mid X_0 = i) = \frac{d}{dz} f_i(t,z)\Big|_{z=1} = i + t \tag{10}$$

in agreement with Lemma 1b.
We also have

$$E(X_t(X_t-1) \mid X_0 = i) = \frac{d^2}{dz^2} f_i(t,z)\Big|_{z=1} = 2t^2 - 2t + (4t-1)i + i^2.$$

Hence

$$E(X_t^2 \mid X_0 = i) = E(X_t(X_t-1) \mid X_0 = i) + E(X_t \mid X_0 = i) = 2t^2 - t + 4ti + i^2.$$

Hence

$$\begin{aligned} Var(X_t \mid X_0 = i) &= E(X_t^2 \mid X_0 = i) - E(X_t \mid X_0 = i)^2 \\ &= 2t^2 - t + 4ti + i^2 - (i+t)^2 \\ &= t^2 + (2i-1)t. \end{aligned} \tag{11}$$

3.3 Explicit Expression for $q_k(t)$

As stated above, Theorem 1 (originally from [27]) gives an explicit expression for SI between two genomes, $\mathcal{G}_0$ and $\mathcal{G}_t$. Nevertheless we could not derive an expression only in terms of the number of events occurred during time t, or alternatively 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 (also in terms of (λt). As here we obtained explicit expression for $p_{i,j}(t)$ we can aim now at expressing q_k.

Lemma 3.

$$q_k(t) = \frac{1}{k} \sum_{\ell=0}^{k-1} \sum_{i=0}^{k-\ell-1} \sum_{j=0}^{k-\ell-1} \frac{(i+j+\ell)!}{i!j!\ell!} t^{i+j} (t+1)^{-i-j-2\ell-1} \left(1-t^2\right)^{\ell}. \tag{12}$$

The full proof of Lemma 3 is deferred to the journal version.

4 Additivity of the SI Measure

Our goal now is to prove the monotonicity of the SI measure for any t and, by Theorem 1, of the expression $h_k(t) = e^{-2t} q_k(t)$ in $t \in [0, \infty)$. In fact we will prove that $q_k(t)$ itself is monotone decreasing, which obviously implies that $h_k(t)$ is monotone decreasing. To do so we first obtain expressions for $q_k'(t)$. As $q_{i,k}(t) = \sum_{j=1}^{k} p_{i,j}(t)$, we get $\frac{dq_{i,k}(t)}{dt} = \sum_{j=1}^{k} \frac{dp_{i,j}(t)}{dt}$.

Lemma 4.

$$\frac{dq_{i,k}(t)}{dt} = k\lambda[p_{i,k+1}(t) - p_{i,k}(t)]. \tag{13}$$

The full proof of Lemma 4 is deferred to the journal version.

Now, as by Lemma 4 we have $q'_{i,k}(t) = k[p_{i,k+1}(t) - p_{i,k}(t)]$, hence

$$q'_k(t) = \frac{1}{k}\sum_{i=1}^{k} q'_{i,k}(t) = \sum_{i=1}^{k}[p_{i,k+1}(t) - p_{i,k}(t)] \tag{14}$$

Using the explicit expressions for $p_{i,j}(t)$, we have, for $i \le k$:

$$p_{i,k}(t) = \frac{1}{(t+1)^{i+k-1}} \cdot \sum_{\ell=1}^{i} \frac{(i+k-\ell-1)!}{(i-\ell)!(k-\ell)!(\ell-1)!} \left(1-t^2\right)^{\ell-1} t^{i+k-2\ell}, \tag{15}$$

and

$$\begin{aligned} p_{i,k+1}(t) &= \frac{1}{(t+1)^{i+k}} \cdot \sum_{\ell=1}^{i} \frac{(i+k-\ell)!}{(i-\ell)!(k+1-\ell)!(\ell-1)!} \left(1-t^2\right)^{\ell-1} t^{i+k+1-2\ell} \\ &= \frac{1}{(t+1)^{i+k-1}} \cdot \sum_{\ell=1}^{i} \frac{(i+k-\ell-1)!}{(i-\ell)!(k-\ell)!(\ell-1)!} \cdot \frac{i+k-\ell}{k+1-\ell} \cdot \frac{t}{t+1} \left(1-t^2\right)^{\ell-1} t^{i+k-2\ell} \end{aligned} \tag{16}$$

hence

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

so that

$$\begin{aligned} q'_k(t) &= \sum_{i=1}^{k}[p_{i,k+1}(t) - p_{i,k}(t)] \\ &= \sum_{i=1}^{k} \frac{1}{(t+1)^{i+k-1}} \cdot \sum_{\ell=1}^{i} \frac{(k+i-\ell-1)!}{(i-\ell)!(k-\ell)!(\ell-1)!} \cdot \left(\frac{k+i-\ell}{k+1-\ell} \cdot \frac{t}{t+1} - 1\right) \left(1-t^2\right)^{\ell-1} t^{i+k-2\ell}. \end{aligned} \tag{17}$$

We would like to prove that $q'_k(t) < 0$ for all $k \ge 1, t > 0$. This is not clear from the above expression. In the next section we prove this by advanced computer algebra tools.

4.1 Computer proof of a double-sum identity

This section is dedicated to the proof of the following identity:

$$q'_k(t) = -\frac{1}{(t+1)^{2k}} \sum_{m=0}^{k-1} \binom{k-1}{m}\binom{k}{m} t^{2m}. \tag{18}$$

By (17) we need to prove that

$$\sum_{i=1}^{k}(t+1)^{k-i}\sum_{\ell=1}^{i}\frac{(i+k-\ell-1)!}{(i-\ell)!(k-\ell)!(\ell-1)!}\cdot\left(1-\frac{i-1}{k+1-\ell}\cdot t\right)\left(1-t^2\right)^{\ell-1}t^{i+k-2\ell}$$
$$=\sum_{m=0}^{k-1}\binom{k-1}{m}\binom{k}{m}t^{2m} \tag{19}$$

The strategy is as follows: we first prove that the right-hand side of (19) satisfies a second-order recurrence in k (Lemma 5), then we derive a recurrence equation for the left-hand side (Lemmas 7 and 8). Since it turns out that these two recurrences are the same, the equality is established by comparing a few initial values. A key component in the proof is Zeilberger's algorithm [34]. It takes as input a parametric sum of the form $F(n) := \sum_k f(n,k)$ where n is a (discrete) parameter and k runs from $-\infty$ to $+\infty$, or between summation bounds that are linear expressions in n (the most common situation is $k = 0, \ldots, n$). Moreover, the summand $f(n,k)$ needs to be hypergeometric in both variables, that means, the quotients $f(n+1,k)/f(n,k)$ and $f(n,k+1)/f(n,k)$ are bivariate rational functions in n and k. As output, Zeilberger's algorithm produces a linear recurrence equation with polynomial coefficients for $F(n)$, i.e., a linear relation of the form $c_d(n)F(n+d) + \ldots + c_1(n)F(n+1) + c_0(n)F(n) = 0$ where the c_i are polynomials in n, that is satisfied for all $n \in \mathbb{N}$.
For our calculations below, we have employed the Mathematica package HolonomicFunctions [21].

Theorem 3. *For all $k \in \mathbb{N}$ and t a parameter, identity (19) holds.*

Before we prove Theorem 3 we state few auxiliary lemmas.

Lemma 5. *The right-hand side of* (19)*, i.e., the expression*

$$R_k(t) := \sum_{m=0}^{k-1}\binom{k-1}{m}\binom{k}{m}t^{2m} \tag{20}$$

satisfies the recurrence

$$(k+2)(2k+1)R_{k+2} - 2\left(2k^2t^2+2k^2+4kt^2+4k+2t^2+1\right)R_{k+1} + k(2k+3)(t-1)^2(t+1)^2R_k = 0$$

for all $k \in \mathbb{N}$.

The full proof of Lemma 5 is deferred to the journal version.

The proof of the following lemma uses the same strategy as the one of Lemma 5.

Lemma 6. *The inner sum of the left-hand side of* (19)*, i.e., the expression*

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

satisfies the following bivariate recurrences:

$$\begin{aligned}
&(k+1)t(i-k)M_{k+1,i} - it(t+1)^2(i-k-t-1)M_{k,i+1} \\
&\qquad + (t-1)(t+1)^2\left(i^2-2ik-it-i+k^2+k\right)M_{k,i} = 0, \\
&(i+1)t(t+1)(i-k)M_{k,i+2} \\
&\qquad + \left(-2i^2t^2+i^2+2ikt^2-2ik-2it^2+i+k^2+kt^2-kt-k\right)M_{k,i+1} \\
&\qquad + i(t-1)t(i-k+1)M_{k,i} = 0.
\end{aligned}$$

Lemma 7. *The left-hand side of* (19) *can be simplified to a single sum, i.e., the following identity holds for all* $k \in \mathbb{N}$*:*

$$\begin{aligned}
\sum_{i=1}^{k} M_{k,i}(t) = L_k(t) := \sum_{\ell=1}^{1+k} &\frac{k(1-t)^{\ell-1}t^{1+2k-2\ell}(1+t)^{\ell-2}(2k-\ell)!}{(1+k-\ell)!(2+k-\ell)!(\ell-1)!} \\
&\times \left((1+k-\ell)(2+k-\ell)+(2+k-\ell)t+\left(1-k^2\right)t^2\right)
\end{aligned}$$

with $M_{k,i}(t)$ *as introduced in Lemma 6.*

The full proof of Lemma 7 is deferred to the journal version.

The proof of the following lemma uses the same strategy as the one of Lemma 5.

Lemma 8. *The sum* $L_k(t)$ *defined in Lemma 7 satisfies the recurrence*

$$(k+2)(2k+1)L_{k+2} - 2\left(2k^2t^2+2k^2+4kt^2+4k+2t^2+1\right)L_{k+1} + k(2k+3)(t-1)^2(t+1)^2L_k = 0$$

for all $k \in \mathbb{N}$.

Proof (Proof of Theorem 3). We have shown that both sides of (19) satisfy the same second-order linear recurrence equation (Lemma 5 and Lemma 8). Since the leading coefficient $(k+2)(2k+1)$ is nonzero for all $k \in \mathbb{N}$, it suffices to verify that (19) holds for $k = 0$ (indeed: both sides evaluate to 0) and for $k = 1$ (indeed: both sides evaluate to 1).

Theorem 3 justifies Eqn. (18), which in turn implies that the function $h_k(t) = \exp(-\lambda t)q_k(t)$ is monotone decreasing with t and thus has an inverse (h_k^{-1}). Moreover, $h_k(t)$ can be exactly calculated (using the explicit expression for $q_k(t)$ given by Eqn. (12)), and so, by Theorem 1, the time separating two sequences of genes involving n genes (where n is large) can be estimated by applying h_k^{-1} to the $\overline{SI}$ for the two gene sequences. Since the expected number of transfer events is additive on the tree (and proportional to t), we conclude the following:

Corollary 2. *The topology of the underlying unrooted tree* T *can be reconstructed in a statistically consistent way from the* $\overline{SI}$ *values by applying the transformation* h_k^{-1}*, followed by a consistent distance-based tree reconstruction method such as Neighbour-Joining (NJ).*

5 Conclusions

In this paper, we have provided an alternative derivation for the system variables of the birth-death formulation of the synteny index (SI) distance measure. The classical approach for this task uses the so-called Karlin-McGregor spectral representation, that is based on a sequence of orthogonal polynomials and a spectral measure [16,17,18,3]. The approach presented here is a self-contained derivation, based on generating functions representation and a subsequent combinatorial treatment, leading to an application of tools from symbolic algebra. Although the biological contribution of this work is seemingly less pronounced, as it merely arrives at the same expressions for the transition probabilities as the traditional approach, we believe that the derivation presented here not only has independent interest for mathematical biology, but may also be key to future rigorous extensions of the Jump model.
One such immediate follow-up extension we see as important is to augment the pure Jump process, with more realistic genome dynamic events such as external gene gain, in which a novel gene is acquired from a different genome, leading to an extension of the gene repertoire of the organism, and events of gene loss. Both these evens potentially cause a divergence in genome content between the analysed genomes, and require special treatment that, based on initial attempts, is non-trivial.

Regarding the mathematical aspect, the symbolic algebra tools as we apply here, have been proved useful in other applications of mathematical biology [5,6] leading to accurate expressions of quantities known to be derived heuristically before. We are hopeful that this new derivation is a basis for the extensions we consider in the future.

References

1. Adato, O., Ninyo, N., Gophna, U., Snir, S.: Detecting horizontal gene transfer between closely related taxa. PLoS computational biology **11**(10), e1004408 (2015)
2. Allen, L.J.: An introduction to stochastic processes with applications to biology. Chapman and Hall/CRC (2010)
3. Anderson, W.J.: Continuous-time Markov chains: An applications-oriented approach. Springer Science & Business Media (2012)
4. Biller, P., Guéguen, L., Tannier, E.: Moments of genome evolution by double cut-and-join. BMC bioinformatics **16**(14), S7 (2015)
5. Chor, B., Hendy, M.D., Snir, S.: Maximum Likelihood Jukes-Cantor Triplets: Analytic Solutions. Mol Biol Evol **23**(3), 626–632 (2006)
6. Chor, B., Khetan, A., Snir, S.: Maximum likelihood on four taxa phylogenetic trees: Analytic solutions. In: Proceedings of the Seventh annual International Conference on Computational Molecular Biology (RECOMB). pp. 76–83. Berlin, Germany (April 2003)
7. Doolittle, W.F.: Phylogenetic classification and the universal tree. Science **284**(5423), 2124–2128 (1999)

8. Felsenstein, J.: Cases in which parsimony or compatibility methods will be positively misleading. Systematic zoology **27**(4), 401–410 (1978)
9. Felsenstein, J.: Evolutionary trees from dna sequences: a maximum likelihood approach. Journal of molecular evolution **17**(6), 368–376 (1981)
10. Fitz Gibbon, S.T., House, C.H.: Whole genome-based phylogenetic analysis of free-living microorganisms. Nucleic acids research **27**(21), 4218–4222 (1999)
11. Hannenhalli, S., Pevzner, P.A.: Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals. vol. 46, pp. 1–27. ACM (1999)
12. Hendy, M.D., Penny, D.: A framework for the quantitative study of evolutionary trees. Systematic zoology **38**(4), 297–309 (1989)
13. Hendy, M.D., Penny, D.: Spectral analysis of phylogenetic data. Journal of classification **10**(1), 5–24 (1993)
14. Hendy, M.D., Penny, D., Steel, M.: A discrete fourier analysis for evolutionary trees. Proceedings of the National Academy of Sciences **91**(8), 3339–3343 (1994)
15. Jaccard, P.: Étude comparative de la distribution florale dans une portion des alpes et des jura. Bull Soc Vaudoise Sci Nat **37**, 547–579 (1901)
16. Karlin, S., McGregor, J.: The classification of birth and death processes. Transactions of the American Mathematical Society **86**(2), 366–400 (1957)
17. Karlin, S., McGregor, J.: A characterization of birth and death processes. Proceedings of the National Academy of Sciences **45**(3), 375–379 (1959)
18. Karlin, S., McGregor, J.L.: The differential equations of birth-and-death processes, and the stieltjes moment problem. Transactions of the American Mathematical Society **85**(2), 489–546 (1957)
19. Katriel, G., Mahanaymi, U., Brezner, S., Kezal, N., Koutschan, C., Zeilberger, D., Steel, M., Snir, S.: Gene transfer-based phylogenetics: Analytical expressions and additivity via birth–death theory. accepted to Systematic Biology (2023)
20. Koonin, E.V., Makarova, K.S., Aravind, L.: Horizontal gene transfer in prokaryotes: quantification and classification. Annual Reviews in Microbiology **55**(1), 709–742 (2001)
21. Koutschan, C.: HolonomicFunctions (user's guide). Tech. Rep. 10-01, RISC Report Series, Johannes Kepler University, Linz, Austria (2010), `http://www.risc.jku.at/research/combinat/software/HolonomicFunctions/`, http://www.risc.jku.at/research/combinat/software/HolonomicFunctions/
22. Miller, S.: The Probability Lifesaver: All the Tools You Need to Understand Chance. Princeton Lifesaver Study Guides, Princeton University Press (2017), `https://books.google.co.il/books?id=VwtHvgAACAAJ`
23. Ochman, H., Lawrence, J.G., Groisman, E.A.: Lateral gene transfer and the nature of bacterial innovation. nature **405**(6784), 299 (2000)
24. Sankoff, D.: Edit distance for genome comparison based on non-local operations. In: Annual Symposium on Combinatorial Pattern Matching. pp. 121–135. Springer (1992)
25. Sankoff, D., Nadeau, J.H.: Conserved synteny as a measure of genomic distance. Discrete applied mathematics **71**(1-3), 247–257 (1996)
26. Serdoz, S., Egri-Nagy, A., Sumner, J., Holland, B.R., Jarvis, P.D., Tanaka, M.M., Francis, A.R.: Maximum likelihood estimates of pairwise rearrangement distances. Journal of theoretical biology **423**, 31–40 (2017)
27. Sevillya, G., Doerr, D., Lerner, Y., Stoye, J., Steel, M., Snir, S.: Horizontal Gene Transfer Phylogenetics: A Random Walk Approach. Molecular Biology and Evolution **37**(5), 1470–1479 (12 2019). https://doi.org/10.1093/molbev/msz302, `https://doi.org/10.1093/molbev/msz302`

28. Shifman, A., Ninyo, N., Gophna, U., Snir, S.: Phylo si: a new genome-wide approach for prokaryotic phylogeny. Nucleic acids research **42**(4), 2391–2404 (2013)
29. Snel, B., Bork, P., Huynen, M.A.: Genome phylogeny based on gene content. Nature genetics **21**(1), 108 (1999)
30. Tekaia, F., Dujon, B.: Pervasiveness of gene conservation and persistence of duplicates in cellular genomes. Journal of molecular evolution **49**(5), 591–600 (1999)
31. Wang, L.S., Warnow, T.: Estimating true evolutionary distances between genomes. In: Proceedings of the thirty-third annual ACM symposium on Theory of computing. pp. 637–646. ACM (2001)
32. Wilf, H.S., Zeilberger, D.: An algorithmic proof theory for hypergeometric (ordinary and "q") multisum/integral identities. Inventiones Mathematicae **108**(1), 575–633 (1992)
33. Yancopoulos, S., Attie, O., Friedberg, R.: Efficient sorting of genomic permutations by translocation, inversion and block interchange. Bioinformatics **21**(16), 3340–3346 (2005)
34. Zeilberger, D.: A fast algorithm for proving terminating hypergeometric identities. Discrete Mathematics **80**(2), 207–211 (1990). https://doi.org/10.1016/0012-365X(90)90120-7
35. Zeilberger, D.: A holonomic systems approach to special functions identities. Journal of Computational and Applied Mathematics **32**(3), 321–368 (1990)