Estimating True Evolutionary Distances under Rearrangements, Duplications, and Losses

Yu Lin[*], Vaibhav Rajan, Krister M. Swenson and Bernard M.E. Moret[*]

Laboratory for Computational Biology and Bioinformatics, Swiss Federal Institute of Technology (EPFL), EPFL-IIS-LCBB, INJ 230, Station 14, CH-1015 Lausanne, Switzerland

Email: yu.lin@epfl.ch; vaibhav.rajan@epfl.ch; krister.swenson@epfl.ch; bernard.moret@epfl.ch;

[*]Corresponding author

Abstract

Background: The rapidly increasing availability of whole-genome sequences has enabled the study of whole-genome evolution. Evolutionary mechanisms based on genome rearrangements have attracted much attention and given rise to many models; somewhat independently, the mechanisms of gene duplication and loss have seen much work. However, the two are not independent and thus require a unified treatment, which remains missing to date. Moreover, existing rearrangement models do not fit the dichotomy between most prokaryotic genomes (one circular chromosome) and most eukaryotic genomes (multiple linear chromosomes).

Results: To handle rearrangements, gene duplications and losses, we propose a new evolutionary model and the corresponding method for estimating true evolutionary distance. Our model, inspired from the DCJ model, is simple and the first to respect the prokaryotic/eukaryotic structural dichotomy. Experimental results on a wide variety of genome structures demonstrate the very high accuracy and robustness of our distance estimator.

Conclusions: We give the first robust, statistically based, estimate of genomic pairwise distances based on rearrangements, duplications and losses, under a model that respects the structural dichotomy between prokaryotic and eukaryotic genomes. Accurate and robust estimates in true evolutionary distances should translate into much better phylogenetic reconstructions as well as more accurate genomic alignments, while our new model of genome rearrangements provides another refinement in simplicity and verisimilitude.

Background

Introduction

Interest in the evolution of genome structure has been growing steadily in the last 10 years, sustained in part by the ever increasing number of sequenced genomes. In particular much work has been done on rearrangements (see, e.g., [1]), using the convention that each chromosome of the genome is represented by an ordered list of identifiers, each identifier referring to a syntenic block or, more commonly, to a member of a gene family. (In the following, we shall use the word "gene" in a broad sense to denote elements of such orderings and refer to such orderings as "gene orders".) Variations in the placement of homologous genes, as well as variations in gene content and multiplicity, among organisms can then be analyzed. Such data is of great interest to evolutionary biologists, but also to comparative genomicists and to any researcher interested in understanding evolutionary changes in pathogens, crop plants, and, more generally, the biome.

The most fundamental task in the analysis of such data is to estimate the amount of evolutionary change between two genomes—that is, to compute a pairwise *evolutionary distance*. The *true distance*, that is, the number of *actual* evolutionary events (rearrangements, duplications, and losses) that took place during the course of evolution, is what we want to obtain, but is not, of course, something that we can compute. Researchers have thus used a two-stage process, in which a well defined measure is first computed (such as an *edit distance*, that is, the *smallest* number of evolutionary events needed to transform one genome into the other), then a statistical model of evolution is used to infer an estimate of the true distance by deriving the effect of a given number of changes in the model on the computed measure and (algebraically or numerically) inverting the derivation to produce a maximum-likelihood estimate of the true distance under the model. This second step is usually called a *distance correction* and has long been used for sequence (DNA) data (see, e.g., [2]) as well as, more recently, for gene-order data, for which see [3–7].

Evolutionary events that affect the gene order of genomes include various rearrangements, which affect only the order, and gene duplications and losses, which affect both the gene content and, indirectly, the order. (Gene insertion, corresponding to lateral gene transfer or neofunctionalization, can be viewed as a special case of duplication.) Rearrangements themselves include inversion, transposition, block exchange, circularization and linearization, all of which act on a single chromosome, and translocation, fusion, and fission, which act on two chromosomes. All of these operations are subsumed in the *double-cut-and-join (DCJ)* [8,9], which has formed the basis for much of the algorithmic research on rearrangements over the last few years, including a statistically based method to estimate the true evolutionary distance between two genomes [7]. DCJ makes two cuts, which can be in the same chromosome or in two different chromosomes, producing four cut ends, then rejoins the four cut ends in any of the three possible ways. The DCJ model, however, is unrealistic in two major respects. First, if the two cuts are in the same chromosome, one of the two nontrivial rejoinings causes a fission, creating a new circular chromosome. However, circular chromosomes do not normally arise in organisms with linear chromosomes, and prokaryotic genomes normally consist of a single circular chromosome. Nor can this form of rejoining be forbidden as, without it, DCJ simply reduces to inversion. Secondly, DCJ is a model of rearrangements: it does not take into account evolutionary events that alter the gene content, such as duplications and losses.

Of these two problems, the first has not been seriously addressed: the model we present here is, to the best of our knowledge, the first model that naturally preserves the dichotomy between prokaryotic and eukaryotic genomes. While gene (or segment) duplications and losses have long been studied by geneticists and molecular biologist, their integration with rearrangements in a unified model has seen relatively little work to date. El-Mabrouk [10] gave an exact algorithm to compute edit distances for inversions and losses and also a heuristic to approximate edit distances for inversions, losses, and nonduplicating insertions (all of her results assume that genes cannot be duplicated). More recently, Yancopoulos and Friedberg [11] gave an algorithm to compute edit distances under deletions, insertions, duplications, and DCJ operations, under the constraint that each deletion can only remove a single gene. These and other approaches targeted the edit distance, not the true evolutionary distance. Swenson *et al.* [12] gave an algorithm to approximate the true evolutionary distance under deletions, insertions, duplications, and inversions for unichromosomal genomes and showed good results under simulations and for small-scale phylogenetic reconstruction. Rearrangements, duplications and losses have also been addressed in the framework of ancestral reconstruction (see, e.g., [13]). All of these approaches have focussed on parsimony criteria and have used preassigned weights for the various operations.

In this paper, we propose a new evolutionary model which respects the dichotomy between prokaryotic and eukaryotic genomes and which takes gene duplications and losses into account. Using this new evolutionary model, we develop a statistically based method to estimate the true evolutionary distance in terms of the actual number of rearrangements, gene duplications, and gene losses. Finally, we provide extensive experimental results on a wide variety of genome structures to illustrate the robustness and high accuracy of our estimator.

Genomes as gene-order data

We denote the tail of a gene g by g^t and its head by g^h. We write $+g$ to indicate an orientation from tail to head ($g^t \rightarrow g^h$), $-g$ otherwise ($g^h \rightarrow g^t$). Two consecutive genes a and b can be connected by one *adjacency* of one of the following four types: $\{a^t, b^t\}$, $\{a^h, b^t\}$, $\{a^t, b^h\}$, and $\{a^h, b^h\}$. If gene c lies at one end of a linear chromosome, then we also have a singleton set, $\{c^t\}$ or $\{c^h\}$, called a *telomere*. A *genome* can then be represented as a multiset of genes together with a multiset of adjacencies and telomeres. For example, the toy genome in Figure 1, composed of one linear chromosome, $(+a,+b,-c,+a,+b,-d,+a)$, and one circular one, $(+e,-f)$, can be represented by the multiset of genes $\{a,a,a,b,b,c,d,e,f\}$ and the multiset of adjacencies and telomeres $\{\{a^t\}, \{a^h,b^t\}, \{b^h,c^h\}, \{c^t,a^h\}, \{a^h,b^t\}, \{b^h,d^h\}, \{d^t,a^h\}, \{a^h\}, \{e^h,f^h\}, \{e^t,f^t\}\}$. Because of the duplicated genes, there is no one-to-one correspondence between genomes and multisets of genes, adjacencies, and telomeres. For example, the genome composed of one linear chromosome, $(+a,+b,-d,+a,+b,-c,+a)$ and one circular one $(+e,-f)$ would have the same multisets of genes, adjacencies and telomeres as that in Figure 1.

Preliminaries on the Evolutionary Model

We use two parameters: the probability of occurrence of a gene duplication, p_d, and the probability of occurrence of a gene loss, p_l; the probability of occurrence of a rearrangement is then just $p_r = 1 - p_d - p_l$. The next event is chosen from the three categories according to these parameters.

For rearrangements, we will select two adjacencies or telomeres with replacement uniformly from the multiset of all adjacencies and telomeres and then decide which rearrangement event we apply. The four cases are as follows.

Select two different adjacencies, or one adjacency and one telomere, in the same chromosome. For example, select two different adjacencies $\{a^h_{i-1},a^t_i\}$ and $\{a^h_j,a^t_{j+1}\}$ on one linear chromosome $C = (a_1 \ldots a_{i-1}a_i \ldots a_j a_{j+1} \ldots a_n)$. Reversing all genes between a_i and a_j yields $(a_1 \ldots a_{i-1} - a_j \ldots - a_i a_{j+1} \ldots a_n)$. Two adjacencies, $\{a^h_{i-1},a^t_i\}$ and $\{a^h_j,a^t_{j+1}\}$, are replaced by two others, $\{a^h_{i-1},a^h_j\}$ and $\{a^t_i,a^t_{j+1}\}$. This operation causes an inversion.

Select two adjacencies or one adjacency and one telomere in two different chromosomes. For example, select two adjacencies, $\{a^h_i,a^t_{i+1}\}$ from one linear chromosome $C = (a_1 \ldots a_i a_{i+1} \ldots a_n)$ and $\{b^h_j,b^t_{j+1}\}$ from another linear chromosome $D = (b_1 \ldots b_j b_{j+1} \ldots b_n)$. Now exchange the two segments between these two chromosomes C and D. There are two possible outcomes, $(a_1 \ldots a_i b_{j+1} \ldots b_n)$ and $(b_1 \ldots b_j a_{i+1} \ldots a_n)$ or $(a_1 \ldots a_i - b_j \ldots - b_1)$ and $(-b_n \ldots - b_{j+1} a_{i+1} \ldots a_n)$. Two adjacencies, $\{a^h_i,a^t_{i+1}\}$ and $\{b^h_j,b^t_{j+1}\}$, are replaced by $\{a^h_i,b^h_{j+1}\}$ and $\{a^t_{i+1},b^t_j\}$ or $\{a^h_i,b^h_j\}$ and $\{a^t_{i+1},b^t_{j+1}\}$. This operation causes a translocation (or, if at least one chromosome is circular, a fusion).

Select the same adjacency twice. For example, select the adjacency $\{a^h_i,a^t_{i+1}\}$ twice from linear chromosome $C = (a_1 \ldots a_i a_{i+1} \ldots a_n)$. Then split C into two new linear chromosomes, $(a_1 \ldots a_i)$ and $(a_{i+1} \ldots a_n)$. The adjacency $\{a^h_i,a^t_{i+1}\}$ is replaced by two telomeres $\{a^h_i\}$ and $\{a^t_{i+1}\}$. This operation causes a fission for a linear chromosome, a linearization for a circular one.

Select two telomeres.[1] For example, select telomeres $\{a^h_i\}$ and $\{b^t_j\}$ from two different linear chromosomes. Then concatenate these two linear chromosomes into a single new chromosome. Two telomeres, $\{a^h_i\}$ and $\{b^t_j\}$, are replaced by two other telomeres, $\{a^h_i$ and $b^t_j\}$. This operation causes a fusion on two linear chromosomes or a circularization on one linear chromosome.

For gene duplication, we uniformly select a position to start duplicating a short segment of chromosomal material and place the new copy to a new position within the genome. We set L_{max} as the maximum number of genes in the duplicated segment and assume that the number of genes in that segment is a uniform random number between 1 and L_{max}. For example, select one segment $a_{i+1} \ldots a_{i+L}$ to duplicate and insert the copy between one adjacency $\{b^h_j,b^t_{j+1}\}$. Such an operation duplicates L genes and $L-1$ adjacencies, removes one adjacency, and adds two new adjacencies; thus genes a_{i+1}, …, a_{i+L-1} and a_{i+L} are added to the multiset of genes, the adjacency $\{b^h_j,b^t_{j+1}\}$ is removed, and $L+1$ new adjacencies, $\{b^h_j,a^t_{i+1}\}$, $\{a^h_{i+1},a^t_{i+2}\}$, …, $\{a^h_{i+L},b^t_{j+1}\}$, are added.

For gene loss, we restrict deletion to genes with at least two copies in the genome and we delete one gene at a time. We uniformly select one gene from the set of all candidate genes and delete it. For example, if we delete gene a_i in the chromosome $(\ldots a_{i-1}a_i a_{i+1} \ldots)$, one copy of a_i is removed from the multiset of genes, while two adjacencies, $\{a^h_{i-1},a^t_i\}$ and $\{a^h_i,a^t_{i+1}\}$, are replaced by one adjacency, $\{a^h_{i-1},a^t_{i+1}\}$.

[1] Selecting one telomere twice is assimilated to selecting both telomeres of the linear chromosome.

Methods

An overview of our technique for estimating the true evolutionary distance

The problem of estimating the true evolutionary distance is defined as follows:

Input: The original genome G and the final genome F.

Output: An estimate of the actual number of evolutionary events that took place in the evolutionary history to transform G into F.

Based on the multisets of genes and of adjacencies and telomeres of G, for any genome G^* of N^* genes and l^* linear chromosomes, we can build the vector $V^* = (NG_1^*, \ldots, NG_C^*, SA_1^*, \ldots, SA_C^*, DA^*, ST^*, DT^*)$, where C is the upper bound for the number of copies of one gene, NG_i^* $(i = 1, \ldots, C)$ is the number of genes with *exactly* i copies in the genome G^*, SA_1^* $(i = 1, \ldots, C)$ is the number of adjacencies with *exactly* i copies in G^* that also appear in G, DA^* is the number of adjacencies in G^* that do not appear in G, ST^* is the number of telomeres in G^* that also appear in G, and DT^* is the number of telomeres in G^* that do not appear in G. We can write

$$\begin{aligned} N^* &= \sum_{i=1}^{C} NG_i^*, \\ N^* &= \sum_{i=1}^{C} SA_1^* + DA^* + ST^* + DT^* - l^*. \end{aligned}$$

Let G^k be the genome obtained from $G = G^0$ by applying k randomly selected evolutionary operations—under our model, the $(i+1)$st evolutionary operation is selected from all possible rearrangements, gene duplications, and gene losses on genome G^i according to the parameters p_d and p_l. We can compute the vector $V_G(G^k) = (NG_1^k, \ldots, NG_C^k, SA_1^k, \ldots, SA_C^k, DA^k, ST^k, DT^k)$ to represent the genome G^k with respect to G.

In the next section, we show that, given G, we can also produce the *estimate* $\widetilde{V_G(G^k)} = (\widetilde{NG_1^k}, \ldots, \widetilde{NG_C^k}, \widetilde{SA_1^k}, \ldots, \widetilde{SA_C^k}, \widetilde{DA^k}, \widetilde{ST^k}, \widetilde{DT^k})$ for the expected vector $E(V_G(G^k))$, for any integer $k > 0$. Our approach for estimating the true evolutionary distance is then to return the integer k that minimizes the 1-norm distance between $\widetilde{V_G(G^k)}$ and $V_G(F)$.

Estimation of the expected vector after some number of random evolutionary events

Given the original genome G, the complete vector for genome G^k is defined as $V_G(G^k) = (NG_1^k, NG_2^k, \ldots, SA_1^k, SA_2^k, \ldots, DA^k, ST^k, DT^k)$, where NG_i^k is the number of genes with exactly i copies in the genome G^k, SA_i^k (shared adjacencies) is the number of adjacencies with exactly i copies in G^k that also appear in G, DA^k (distinct adjacencies) is the number of adjacencies in G^k that do not appear in G, ST^k (shared telomeres) is the number of telomeres in G^k that also appear in G, and DT^k (distinct telomeres) is the number of telomeres in G^k that do not appear in G.

Assume the original genome G has N genes, where each gene has at most $C = O(1)$ copies, and l linear chromosomes, with $l = O(1)$. We thus ignore items NG_i^k and SA_i^k for $(i > C)$. The initial vector $V_G(G^0)$ is then $(NG_1^0, NG_2^0, \ldots, NG_C^0, SA_1^0, SA_2^0, \ldots, SA_C^0, DA^0, ST^0, DT^0)$, where NG_i^0 is the number of genes with exactly i copies, SA_i^0 is the number of adjacencies with exactly i copies, $DA^0 = 0$, $ST^0 = 2l$, and $DT^0 = 0$. We now show how to update this vector under rearrangements, gene duplications and gene losses, respectively.

Rearrangements

We select two adjacencies or telomeres uniformly, with replacement, from the multiset of all adjacencies or telomeres.

Lemma 1 *Assume all genomes have $O(1)$ linear chromosomes, each gene has at most $C = O(1)$ copies, and $V_G(G^k) = (NG_1^k, \ldots, NG_C^k, SA_1^k, \ldots, SA_C^k, DA^k, ST^k, DT^k)$ represents the current genome G^k based on the original genome G. For conciseness, write $N^k = \sum_{i=1}^{C} NG_1^i$ (the total number of genes) and $l^k = (ST^k + DT^k)/2$ (the number of linear chromosomes). Then we can write the expected vector for G^{k+1} after one rearrangement operation: $E(V_G(G^{k+1})) = (NG_1^{k+1}, \ldots, NG_C^{k+1}, SA_1^{k+1}, \ldots, SA_C^{k+1}, DA^{k+1}, ST^{k+1}, DT^{k+1})$ where we have*

$$\begin{aligned} NG_i^{k+1} &= NG_i^k,\ \ i = 1, 2, \ldots, C \\ SA_i^{k+1} &= SA_i^k - \frac{2i(SA_i^k - SA_{i+1}^k)}{N^k + l^k} + O(\frac{1}{N^k}),\ \ i = 1, 2, \ldots, C-1 \\ SA_C^{k+1} &= SA_C^k - \frac{2C(SA_C^k)}{N^k + l^k} + O(\frac{1}{N^k}), \\ DA^{k+1} &= DA^k + \frac{2(\sum_{i=1}^{C} SA_i^k)}{N^k + l^k} + O(\frac{1}{N^k}), \\ ST^{k+1} &= ST^k - \frac{2ST^k}{N^k + l^k} + O(\frac{1}{N^k}) \\ DT^{k+1} &= DT^k + \frac{2ST^k}{N^k + l^k} + O(\frac{1}{N^k}). \end{aligned}$$

Proof In our evolutionary model, each rearrangement operation replaces old adjacencies or telomeres with new ones. Obviously, any rearrangement operation will not change the gene content, so $NG_i^{k+1} (i = 1, 2, \ldots, C)$ will be the same.

We first ignore the adjacencies or telomeres in the original genome G created after a rearrangement event. Remember two adjacencies or telomeres are selected with replacement uniformly from the multiset of all adjacencies and telomeres, and the number of all adjacencies or telomeres for genome G^k is (N^k+l^k). For SA_i^k adjacencies with exactly i copies in G^k which also appear in G, the probability that one adjacency is selected once is $\frac{2SA_i^k(N^k+l^k-SA_i^k)}{(N^k+l^k)^2}$, the probability that two different adjacencies are selected is $\frac{SA_i^k(SA_i^k-i)}{(N^k+l^k)^2}$, the probability that same adjacencies at two different sites are selected is $\frac{(i-1)SA_i^k}{(N^k+l^k)^2}$, and the probability that same adjacency at the same site is selected twice is $\frac{SA_i^k}{(N^k+l^k)^2}$. Ignoring the newly created adjacencies or telomeres in the original genome G, with probability $\frac{2SA_i^k(N^k+l^k-SA_i^k)+iSA_i^k}{(N^k+l^k)^2}$, the number of adjacencies with exactly i copies decreases by i, and, with probability $\frac{SA_i^k(SA_i^k-i)}{(N^k+l^k)^2}$, the number of adjacencies with exactly i copies decreases by $2i$. With probability $\frac{2SA_i^k(N^k+l^k-SA_i^k)+SA_i^k}{(N^k+l^k)^2}$, the number of adjacencies with exactly $(i-1)$ copies increases by $(i-1)$, with probability $\frac{SA_i^k(SA_i^k-i)}{(N^k+l^k)^2}$, the number of adjacencies with exactly $(i-1)$ copies decreases by $2(i-1)$, and, with probability $\frac{(i-1)SA_i^k}{(N^k+l^k)^2}$, the number of adjacencies with exactly $(i-2)$ copies increases by $(i-2)$. Considering $i=1,2,\ldots,C$ and $C=O(1)$, we have

$$\begin{aligned}
SA_i^{k+1} &= SA_i^k - \frac{2i(SA_i^k - SA_{i+1}^k)}{N^k+l^k},\ \ i=1,2,\ldots,C-1\\
SA_C^{k+1} &= SA_C^k - \frac{2C(SA_C^k)}{N^k+l^k},\\
DA^{k+1} &= DA^k + \frac{2(\sum_{i=1}^C SA_i^k)}{N^k+l^k}.
\end{aligned}$$

Now, we show that the correction for our ignoring adjacencies or telomeres after a rearrangement event is $O(\frac{1}{N^k})$ for each item. Consider any adjacency (a,b) in G: we might recover it if we select two adjacencies or telomeres containing two genes a and b. Since each gene has at most C copies in the genome, there are at most C^2 pairs of adjacencies or telomeres that may lead to recovery of the adjacency (a,b). So, with probability at most $\frac{C^2}{(N^k+l^k)^2}$, one specific adjacency in G might be created by the rearrangement. Summing up all the $N-l$ adjacencies in G, we see that the correction for ignoring the newly created adjacencies or telomeres in G is $O(\frac{1}{N^k})$.

Similarly, we can get $ST^{k+1}=ST^k-\frac{2ST^k}{N^k+l^k}+O(\frac{1}{N^k})$ and $DT^{k+1}=DT^k+\frac{2ST^k}{N^k+l^k}+O(\frac{1}{N^k})$.

Gene duplication

We select uniformly at random an integer between 1 and L_{max} (the maximum number of genes in the duplication segment), then select uniformly at random a position in the genome where to start the duplication, then insert the copy at another position selected uniformly at random.

Lemma 2 *Assume all genomes have $O(1)$ linear chromosomes, each gene has at most $C=O(1)$ copies, no two same genes or adjacencies are within the segment to be duplicated, and $V_G(G^k)=(NG_1^k,\ldots,NG_C^k,SA_1^k,\ldots,SA_C^k,DA^k,ST^k,DT^k)$ represents the current genome G^k based on the original genome G. For conciseness, write $N^k=\sum_{i=1}^C NG_1^i$ (the total number of genes), $l^k=(ST^k+DT^k)/2$ (the number of linear chromosomes) and $L=(L_{max}+1)/2$ (the average number of genes in a duplication segment). Then we approximate the expected vector for G^{k+1} after one duplication operation with $E(V_G(G^{k+1}))=(NG_1^{k+1},\ldots,NG_C^{k+1},SA_1^{k+1},\ldots,SA_C^{k+1},DA^{k+1},ST^{k+1},DT^{k+1})$ where we have*

$$\begin{aligned}
NG_1^{k+1} &= NG_1^k - \frac{L(NG_1^k)}{N^k},\\
NG_i^{k+1} &= NG_i^k + \frac{iL(NG_{i-1}^k - NG_i^k)}{N^k},\ \ i=2,\ldots,C-1\\
NG_C^{k+1} &= NG_C^k + \frac{CL(NG_{C-1}^k)+L(NG_C^k)}{N^k},\\
SA_1^{k+1} &= SA_1^k - \frac{(L-1)SA_1^k}{N^k-l^k} - \frac{SA_1^k-SA_2^k}{N^k+l^k} + O(\frac{1}{N^k}),\\
SA_i^{k+1} &= SA_i^k + \frac{i(L-1)(SA_{i-1}^k - SA_i^k)}{N^k-l^k} - \frac{i(SA_i^k - SA_{i+1}^k)}{N^k+l^k}\\
&\quad +O(\frac{1}{N^k}), i=2,\ldots,C-1\\
SA_C^{k+1} &= SA_C^k + \frac{C(L-1)SA_{C-1}^k+(L-1)SA_C^k}{N^k-l^k}\\
&\quad -\frac{C(SA_C^k)}{N^k+l^k} + O(\frac{1}{N^k}),\\
DA^{k+1} &= DA^k + \frac{(L-1)DA^k}{N^k-l^k} + \frac{\sum_{i=1}^C SA_i^k + DA^k}{N^k+l^k} + O(\frac{1}{N^k}),\\
ST^{k+1} &= ST^k - \frac{ST^k}{N^k+l^k} + O(\frac{1}{N^k})\\
DT^{k+1} &= DT^k + \frac{ST^k}{N^k+l^k} + O(\frac{1}{N^k}).
\end{aligned}$$

Proof In our model, we uniformly select a position to start duplicating L genes and transpose it to one new uniformly chosen position within the genome. The expected number of genes or adjacencies with exactly i copies within the duplication segment is $L(NG_i^k)/N^k$ or $(L-1)SA_i^k/(N^k-l^k)$. The probability that the placement of the duplicated segment breaks one adjacency in SA_i^k is $SA_i^k/(N^k+l^k)$.

We again first ignore the adjacencies or telomeres in the original genome G created after a duplication event. Since we assume that no two genes or adjacencies are same within the duplication segment, we have

$$\begin{aligned}
NG_1^{k+1} &= NG_1^k - \frac{L(NG_1^k)}{N^k}, \\
NG_i^{k+1} &= NG_i^k + \frac{iL(NG_{i-1}^k - NG_i^k)}{N^k}, \ \ i = 2,\dots,C-1 \\
NG_C^{k+1} &= NG_C^k + \frac{CL(NG_{C-1}^k) + L(NG_C^k)}{N^k}, \\
SA_1^{k+1} &= SA_1^k - \frac{(L-1)SA_1^k}{N^k - l^k} - \frac{SA_1^k - SA_2^k}{N^k + l^k}, \\
SA_i^{k+1} &= SA_i^k + \frac{i(L-1)(SA_{i-1}^k - SA_i^k)}{N^k - l^k} - \frac{i(SA_i^k - SA_{i+1}^k)}{N^k + l^k}, \\
& \quad i = 2,\dots,C-1 \\
SA_C^{k+1} &= SA_C^k + \frac{C(L-1)SA_{C-1}^k + (L-1)SA_C^k}{N^k - l^k} - \frac{C(SA_C^k)}{N^k + l^k}. \\
DA^{k+1} &= DA^k + \frac{(L-1)DA^k}{N^k - l^k} + \frac{\sum_{i=1}^C SA_i^k + DA^k}{N^k + l^k}.
\end{aligned}$$

Now, we show that the correction for our ignoring adjacencies or telomeres after a duplication event is $O(\frac{1}{N^k})$ to each item SA_i^{k+1}. Consider any adjacency (a,b) in G: we might recover it if we move gene a next to gene b after the duplication. Since each gene has at most C copies in the genome, there are at most $2LC^2$ possibly duplication operations to recover that adjacency (a,b). There are altogether $O(L(N^k + l^k)^2)$ different duplication operations. So, with probability $O(\frac{1}{(N^k+l^k)^2})$, one specific adjacency in G might be created by the duplication event. Summing up all the $N - l$ adjacencies in G, we see that the correction for ignoring the newly created adjacencies or telomeres in G is $O(\frac{1}{N^k})$.

Similarly, we can get $ST^{k+1} = ST^k - \frac{ST^k}{N^k+l^k} + O(\frac{1}{N^k})$ and $DT^{k+1} = DT^k + \frac{ST^k}{N^k+l^k} + O(\frac{1}{N^k})$.

Gene loss

We uniformly select one gene with at least two copies and delete it.

Lemma 3 *Assume each gene has at most $C = O(1)$ copies and $V_G(G^k) = (NG_1^k, NG_2^k, \dots, NG_C^k, SA_1^k, SA_2^k, \dots, SA_C^k, DA^k, ST^k, DT^k)$ represents the current genome G^k based on the original genome G. For conciseness, write $N^k = \sum_{i=1}^C NG_i^k$ (the total number of genes) and $l^k = (ST^k + DT^k)/2$ (the number of linear chromosomes). Then we can write the expected vector for G^{k+1} after one rearrangement operation as $E(V_G(G^{k+1})) = (NG_1^{k+1}, \dots, NG_C^{k+1}, SA_1^{k+1}, \dots, SA_C^{k+1}, DA^{k+1}, ST^{k+1}, DT^{k+1})$, where we have*

$$\begin{aligned}
NG_1^{k+1} &= NG_1^k + \frac{NG_2^k}{N^k - NG_1^k}, \\
NG_i^{k+1} &= NG_i^k - \frac{i(NG_i^k - NG_{i+1}^k)}{N^k - NG_1^k}, i = 2,\dots,C-1 \\
NG_C^{k+1} &= NG_C^k - \frac{C(NG_C^k)}{N^k - NG_1^k}.
\end{aligned}$$

Proof In our model of gene loss, one gene with at least two copies is uniformly selected. The number of all possible genes to be deleted is $N^k - NG_1^k$. For NG_i^k $(i > 1)$ genes with exactly i copies in G^k, the probability that one of them is selected and deleted is $\frac{NG_i^k}{N^k - NG_1^k}$. So with probability $\frac{NG_i^k}{N^k - NG_1^k}$, the number of genes with exactly i copies decreases by i and the number of genes with exactly $(i-1)$ copies increases by $(i-1)$.

We ignore the adjacencies or telomeres in the original genome G to be created after one gene loss. For SA_i^k $(i > 2)$ adjacencies with exactly i copies in G^k which also appears in G, it is difficult to compute the number $f_i(del_j)$ of such adjacencies that each single deletion del_j $(j = 1,\dots,N^k - NG_1^k)$ would affect. But we know that each adjacency with exactly i $(i > 2)$ copies must relate to two genes with more than 2 copies, so we have $\sum_{j=1}^{N^k - NG_1^k} f_i(del_j) = 2SA_i^k$. Considering $i = 2,\dots,C$ and $C = O(1)$, we have

$$\begin{aligned}
SA_i^{k+1} &= SA_i^k - \frac{2i(SA_i^k - SA_{i+1}^k)}{N^k - NG_1^k}, \ \ i = 2,\dots,C-1 \\
SA_C^{k+1} &= SA_C^k - \frac{2C(SA_C^k)}{N^k - NG_1^k}.
\end{aligned}$$

For SA_1^k adjacencies with exactly 1 copy in G^k that also appears in G, it is also difficult to compute the number $f_1(del_j)$ of such adjacencies that each single deletion del_j $(j = N^k - NG_1^k)$ would affect. Assume $DSA_1^k (= \sum_{j=1}^{N^k - NG_1^k} f_1(del_j))$ is the count of genes with at least two copies but related to those adjacencies with exactly 1 copy in G^k that also appear in G. We consider the effect of rearrangements, gene duplications and losses, and we approximate as follows:

$$\begin{aligned}
DSA_1^{k+1} &= DSA_1^k + p_r \frac{2(2SA_2^k - DSA_1^k)}{N^k + l^k} \\
&\quad + p_d \left(\frac{2SA_1^k - 2DSA_1^k + 2SA_2^k}{N^k + l^k} - \frac{(L-1)DSA_1^k}{N^k - l^k} \right) \\
&\quad + p_l \frac{2SA_2^k - DSA_1^k(1 + NG_2^k/(N^k - NG_1^k))}{N^k - NG_1^k}, \\
SA_1^{k+1} &= SA_1^k - p_l \frac{DSA_1^k - 2SA_2^k}{N^k - NG_1^k}.
\end{aligned}$$

For telomeres, we simply assume $ST^{k+1} = ST^k$ and $DT^{k+1} = DT^k$.

Finally, we also approximate the number of adjacencies RSA^{k+1} that we could thus ignore under rearrangements, gene duplications, and gene losses, and distribute it to the correction of SA_i^k as follows:

$$\begin{aligned} RSA^{k+1} &= (p_r + \frac{1}{2}p_d)(N-l)(N^k/N)^2/(N^k+l^k)^2 \\ SA_i^{k+1} &= SA_i^k + RSA^{k+1}SA_i^k/(N^k - l^k - DA^k), \\ & \quad i = 1, \ldots, C-1. \end{aligned}$$

Now, given G^0, we estimate $E(V_G(G^k))$ for $k > 0$ by iterating k times the above formulas (using with p_d and p_l); at every step we identify $E(V_G(G^{k-1}))$ with the actual vector $V_G(G^{k-1})$.

Corollary 1 *The estimated vector* $\widetilde{V_G(G^i)} = (\widetilde{NG^i_1}, \ldots, \widetilde{NG^i_C}, \widetilde{SA^i_1}, \ldots, \widetilde{SA^i_C}, \widetilde{DA^i}, \widetilde{ST^i}, \widetilde{DT^i})$ *for all integers* i $(0 \leq i \leq k)$ *can be computed in* $O(kC)$ *time.*

Experimental Results

We now present experimental results on the accuracy of our estimation of the expected vector after a given number of random evolutionary events and on the quality of our estimator for the true evolutionary distance (in terms of the actual number of evolutionary events). Our experiments all start with one genome with no duplicated genes and some chosen number of linear and circular chromosomes of various sizes. We first apply some number (usually 10) of duplication events ($L_{max} = 10$ in all cases) to generate the original genome G with some initial duplicated genes. Then this genome is subjected to a prescribed number k of evolutionary events chosen according to p_d and p_l to obtain a final genome G^k. We vary k from 0 to twice the number of genes. We ran tests on any types of initial genomes designed to resemble actual organismal genomes; we tested different choices of parameters on different genomes; and in each case we generated $10,000$ runs to obtain a tight estimate of variance.

We compute the vector representations for all intermediate genomes and then use our method to estimate the evolutionary distance. Due to space limitations, we present results on just three initial genomes: $25,000$ genes and 25 linear chromosomes ($p_d = 0.05, p_l = 0.15$); $10,000$ genes and 5 linear chromosomes ($p_d = 0.1, p_l = 0.2$); and $1,000$ genes and 1 circular chromosome ($p_d = 0.2, p_l = 0.6$). The first two examples match large and smaller metazoan genomes, the last matches a small bacterial genome.

Accuracy of the expected vector after k random evolutionary events

We study the behavior of our estimator $\widetilde{V_G(G^k)}$ by comparing its prediction to the sample mean for $V_G(G^k)$, as computed from our $10,000$ trials. In all of our experiments, we find that $\widetilde{V_G(G^k)}$ is very close to the sample mean for $V_G(G^k)$. Figure 2 shows the values in the vector as a function of the actual number of evolutionary events. SA_3^k and NA_3^k represent the number of adjacencies and genes with at least 3 copies in the original genome G, respectively. The figure shows that our estimation and the sample mean for $V_G(G^k)$ are always very close.

Accuracy of the estimation of the actual number of evolutionary events

We want to study the accuracy of our estimator for the actual number of evolutionary events; in order to do that, we create simulations with controlled numbers of evolutionary events and set up a threshold for correction in the estimation procedure. Specifically, we vary the actual number of evolutionary events from 0 to twice the number of genes in the original genome and we set 4 times the number of genes as an upper limit on the maximum number of evolutionary events. C is set to 10. Thus our estimated number k is chosen to minimize $|\widetilde{V_G(G^k)} - V_G(F)|_1$, the 1-norm distance between $\widetilde{V_G(G^k)}$ and $V_G(F)$.

Figure 3 shows the mean and standard deviation for the actual number of evolutionary events estimated by our approach. Our approach provides accurate estimates, with very small variance.

We also study the mean absolute difference between the actual number of evolutionary events and our estimator, shown in Figure 4.

Table 1 shows that the estimates are quite accurate up to very large numbers of events. Rearrangements, gene duplications, and gene losses fall under the category of "rare genomic events" (in the terminology of [14]), yet our estimator works well even for numbers that would instead indicate common events.

Robustness to unknown model parameters

Up to now we have fixed p_d and p_l. We now consider the case in which these parameters are unknown—clearly the more common case in practice. We generate $10,000$ cases with randomly chosen parameters p_d and p_l (at 1% resolution, $p_d < 4p_l$) and with actual numbers of evolutionary events varying from 0 to twice the number of genes, setting an upper limit of 4 times the number genes for the maximum number of evolutionary events.

Given the original genome, our estimated vector $\widetilde{V_G(G^i)}$ is in fact a function of i, p_d, and p_l. We enumerate all possible values for p_d and p_l (at 1% resolution, $p_d < 4p_l$). For each different pair of parameters p_d and p_l, we compute all $\widetilde{V_G(G^i)}$ (i from 0 to 4 times the number of genes, C is set to 10). Our estimated number k is still chosen to minimize $|\widetilde{V_G(G^k)} - V_G(F)|_1$, the 1-norm distance between $\widetilde{V_G(G^k)}$ and $V_G(F)$.

Figure 5 shows the comparison of our estimates to the actual number of evolutionary events. Our approach still provides accurate estimates in absence of known values for p_d and p_l and thus is quite robust. The mean absolute difference between the actual number of evolutionary events and our estimator becomes larger, especially when there are few common adjacencies left between the original and final genomes. (The duplications and losses may also partially cancel each other.)

Discussion and Conclusions

We propose a new evolutionary model for rearrangements, gene duplications and losses, and a corresponding method for estimating true evolutionary distance. The model is, to our knowledge, the first to preserve the structural dichotomy in genomic organization between most prokaryotes and most eukaryotes, and one of the few to unite rearrangements, duplications, and losses. Experimental results on a wide variety of genome structures exemplify the high accuracy and robustness of our estimator. This large gain in accuracy should translate into much better phylogenetic reconstructions as well as more accurate genomic alignments.

Competing interest

None to declare.

References

1. Fertin G, Labarre A, Rusu I, Tannier E, Vialette S: *Combinatorics of Genome Rearrangements*. MIT Press 2009.

2. Swofford D, Olsen G, Waddell P, Hillis D: **Phylogenetic Inference**. In *Molecular Systematics*. Edited by Hillis D, Mable B, Moritz C, Sinauer Assoc., Sunderland, MA 1996:407–514.

3. Moret B, Tang J, Wang LS, Warnow T: **Steps toward accurate reconstructions of phylogenies from gene-order data**. *J. Comput. Syst. Sci.* 2002, **65**(3):508–525.

4. Sankoff D, Blanchette M: **Probability models for genome rearrangement and linear invariants for phylogenetic inference**. In *Proc. 3rd Conf. Research Comput. Mol. Biol. (RECOMB'99)*, ACM Press, New York 1999:302–309.

5. Wang LS: **Exact-IEBP: a new technique for estimating evolutionary distances between whole genomes**. In *Proc. 33rd ACM Symp. Theory of Comput. (STOC'01)*, ACM Press, New York 2001:637–646.

6. Wang LS, Warnow T: **Estimating true evolutionary distances between genomes**. In *Proc. 1st Workshop Algs. in Bioinformatics (WABI'01)*, no. 2149 in Lecture Notes in Comp. Sci., Springer Verlag, Berlin 2001:176–190.

7. Lin Y, Moret B: **Estimating true evolutionary distances under the DCJ model**. In *Proc. 16th Conf. Intelligent Systems for Mol. Biol. (ISMB'08), Volume 24(13) of* Bioinformatics 2008:i114–i122.

8. Yancopoulos S, Attie O, Friedberg R: **Efficient sorting of genomic permutations by translocation, inversion and block interchange**. *Bioinformatics* 2005, **21**(16):3340–3346.

9. Bergeron A, Mixtacki J, Stoye J: **A unifying view of genome rearrangements**. In *Proc. 6th Workshop Algs. in Bioinformatics (WABI'06)*, no. 4175 in Lecture Notes in Comp. Sci., Springer Verlag, Berlin 2006:163–173.

10. El-Mabrouk N: **Genome rearrangement by reversals and insertions/deletions of contiguous segments**. In *Proc. 11th Symp. Combin. Pattern Matching (CPM'00), Volume 1848 of* Lecture Notes in Comp. Sci., Springer Verlag, Berlin 2000:222–234.

11. Yancopoulos S, Friedberg R: **Sorting genomes with insertions, deletions and duplications by DCJ**. In *Proc. 6th RECOMB Workshop Comp. Genomics (RECOMBCG'08), Volume 5267 of* Lecture Notes in Comp. Sci., Springer Verlag, Berlin 2008:170–183.

12. Swenson K, Marron M, Earnest-DeYoung J, Moret B: **Approximating the true evolutionary distance between two genomes**. In *Proc. 7th SIAM Workshop Alg. Engin. & Experiments (ALENEX'05)*, SIAM Press, Philadelphia 2005:121–129.

13. Ouangraoua A, Boyer F, McPherson A, Tannier E, Chauve C: **Prediction of contiguous regions in the amniote ancestral genome**. In *Proc. 5th Int'l Symp. on Bioinformatics Research and Applications, (ISBRA'09), Volume 5542 of* Lecture Notes in Comp. Sci., Springer Verlag, Berlin 2009:173–185.

14. Rokas A, Holland P: **Rare genomic changes as a tool for phylogenetics**. *Trends in Ecol. and Evol.* 2000, **15**:454–459.

Figures

Figure 1 - A very small genome G

Figure 2 - The vector values as a function of the actual number of evolutionary events. $a)$ the color and shape code for the values, $b)$ Gene $\# = 1000$, Linear Chromosome $\# = 0$, Circular Chromosome $\# = 1$, $c)$ Gene $\# = 10000$, Linear Chromosome $\# = 10$, Circular Chromosome $\# = 0$, $d)$ Gene $\# = 25000$, Linear Chromosome $\# = 25$, Circular Chromosome $\# = 0$.

Figure 3 - Mean (indicated by $\times$) and standard deviation (indicated by vertical bar) plots for the actual number of evolutionary events (x axis) vs. our estimator (y axis). $a)$ Gene $\# = 1000$, Linear Chromosome $\# = 0$, Circular Chromosome $\# = 1$, $b)$ Gene $\# = 10000$, Linear Chromosome $\# = 10$, Circular Chromosome $\# = 0$, $c)$ Gene $\# = 25000$, Linear Chromosome $\# = 25$, Circular Chromosome $\# = 0$.

Figure 4 - The mean absolute difference between actual number of different evolutionary events and our estimation as a function of the actual number of evolutionary events (o: Rearrangements, $+$: Duplications, $\times$: Losses). $a)$ Gene $\# = 1000$, Linear Chromosome $\# = 0$, Circular Chromosome $\# = 1$, $b)$ Gene $\# = 10000$, Linear Chromosome $\# = 10$, Circular Chromosome $\# = 0$, $c)$ Gene $\# = 25000$, Linear Chromosome $\# = 25$, Circular Chromosome $\# = 0$.

Figure 5 - $a)$: Mean (indicated by $\times$) and standard deviation (indicated by vertical bar) plots for the actual number of evolutionary events vs. our estimator (Gene $\# = 1000$, Linear Chromosome $\# = 0$, Circular Chromosome $\# = 1$). $b)$: The mean absolute difference between actual number of different evolutionary events and our estimation (o: Rearrangements, $+$: Duplications, $\times$: Losses).

Tables

Table 1 - Relative error of our estimator as a function of the actual number of evolutionary events

# genes	actual number of evolutionary events					
	# genes ×1			# genes ×2		
	Rearrangements	Duplications	Losses	Rearrangements	Duplications	Losses
1000	7.4 %	3.4 %	7.4 %	6.9 %	3.4 %	6.9 %
10,000	1.7 %	1.4 %	2.7 %	2.6 %	1.4 %	3.1 %
25,000	1.3 %	1.5 %	2.0 %	2.6 %	1.5 %	2.9 %