A New Genomic Evolutionary Model for Rearrangements, Duplications, and Losses that Applies across Eukaryotes and Prokaryotes

Yu Lin and Bernard M.E. Moret

Laboratory for Computational Biology and Bioinformatics,
Swiss Federal Institute of Technology (EPFL),
EPFL-IC-LCBB, INJ 230, Station 14, CH-1015 Lausanne, Switzerland
Email: {yu.lin,bernard.moret}@epfl.ch

Abstract.
Background: Genomic rearrangements have been studied since the beginnings of modern genetics and models for such rearrangements have been the subject of many papers over the last 10 years. However, none of the extant models can predict the evolution of genomic organization into circular unichromosomal genomes (as in most prokaryotes) and linear multichromosomal genomes (as in most eukaryotes). Very few of these models support gene duplications and losses—yet these events may be more common in evolutionary history than rearrangements and themselves cause apparent rearrangements.
Results: We propose a new evolutionary model that integrates gene duplications and losses with genome rearrangements and that leads to genomes with either one (or a very few) circular chromosome or a collection of linear chromosomes. Moreover, our model predictions fit observations about the evolution of gene family sizes as well as existing predictions about the growth in the number of chromosomes in eukaryotic genomes. Finally, our model is based on the existing inversion/translocation models and inherits their linear-time algorithm for pairwise distance computation.

1 Introduction

Genomic rearrangements have been studied since the beginnings of modern genetics (starting in the 1920s with the classic work of Sturtevant and Dobzhansky [18, 19]) and models for such rearrangements have been the subject of many papers over the last 20 years (for a review, see [4]). However, none of the extant models predicts the evolution of genomic organization into circular unichromosomal genomes (as in most prokaryotes) and linear multichromosomal genomes (as in most eukaryotes). In addition, hardly any of these models support gene duplications and losses alongside rearrangements; yet duplications and losses may be more common in evolutionary history than rearrangements and, moreover, they themselves cause apparent rearrangements.

In this paper, we propose a new evolutionary model, based on the classical inversion/translocation (HP) [6] and double-cut-and-join (DCJ) [2, 22] models,

that integrates gene duplications and losses with genome rearrangements and that leads to genomes with either a single (or a very few) circular chromosome or a collection of linear chromosomes. Moreover, our model predictions fit observations (as presented by Lynch [15]) about the evolution of gene family sizes, as well as existing predictions (by Imai's group [11]) about the growth in the number of chromosomes in eukaryotic genomes. Finally, our model inherits the algorithmic results developped for previous models, such as a linear-time distance computation [1–3].

2 Background

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 of a gene duplicate, can be viewed as a special case of duplication.)

Rearrangements themselves include inversions, transpositions, block exchanges, circularizations, and linearizations, all of which act on a single chromosome, and translocations, fusions, and fissions, which act on two chromosomes. These operations are subsumed in the *double-cut-and-join (DCJ)* [2, 22], which has formed the basis for much algorithmic research on rearrangements over the last few years. A DCJ operation 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 is more general than the HP model, because it applies equally well to circular and linear chromosomes. However, the DCJ model still falls short in two 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, while most prokaryotic genomes consist of a single circular chromosome. This unrealistic operation can be corrected by forcing reabsorption of circular intermediates right after their introduction [22]. But this additional constraint creates dependencies among blocks of steps, which introduces difficulties in the estimation of the true distances (see [**?**]). Secondly, DCJ is a model of rearrangements: it does not take into account evolutionary events that alter the gene content and also, indirectly, the gene order, such as duplications and losses.

Genome evolution appears driven by very general mechanisms. For instance, for a wide variety of genomic properties, the number of families of a given size usually declines with the size of the family, following some asymptotic power law, the most common family size being one. Such scaling holds for gene families [7], protein folds and families (encoded in genomes) [12], and pseudogene families and pseudomotifs [14]. Several evolutionary models [5, 7, 16, 21], all

based on gene duplication, have been proposed to explain the observed biological data. More recently, Lynch [15] observed that the frequency distributions of family sizes observed in different species tend to bow downward rather than obey a power law. He gave a simple birth/death model to account for these observations. In this model, each gene (including duplicated ones) has pre-generation probability D (for duplication) of giving rise to a new copy, such that the average birth rate of a family of x members is Dx. The model also assumes that the presence of at least one member of the gene family is essential (i.e., complete loss of the gene family is not possible), but all excess copies have a probability L (for loss) of being eliminated. With D/L ratios consistent with actual estimates for eukaryotic genes, the equilibrium probability distribution of gene family sizes is close to the observations.

Our model of genomic evolution includes all of the operations from DCJ model, except the aforementioned operation that creates circular intermediates, and hence subsumes the HP model [6]; it also takes gene duplications and losses into account, all in a single step. The new evolutionary model respects the distinction between prokaryotic and eukaryotic genomes and also agrees with current predictions about genomic evolution, such as the distribution of sizes of gene families and the number of chromosomes in eukaryotic genomes. In earlier work, we described a method for estimating precisely true evolutionary distance between two genomes under this model using the independence among steps [13].

2.1 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 have a corresponding 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 composed of one linear chromosome, $(+a,+b,\text{-}c,+a,+b,\text{-}d,+a)$, and one circular one, $(+e,\text{-}f)$, 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 the linear chromosome $(+a, +b, \text{-}d, +a, +b, \text{-}c, +a)$ and the circular one $(+e, \text{-}f)$, would have the same multisets of genes, adjacencies and telomeres.

2.2 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

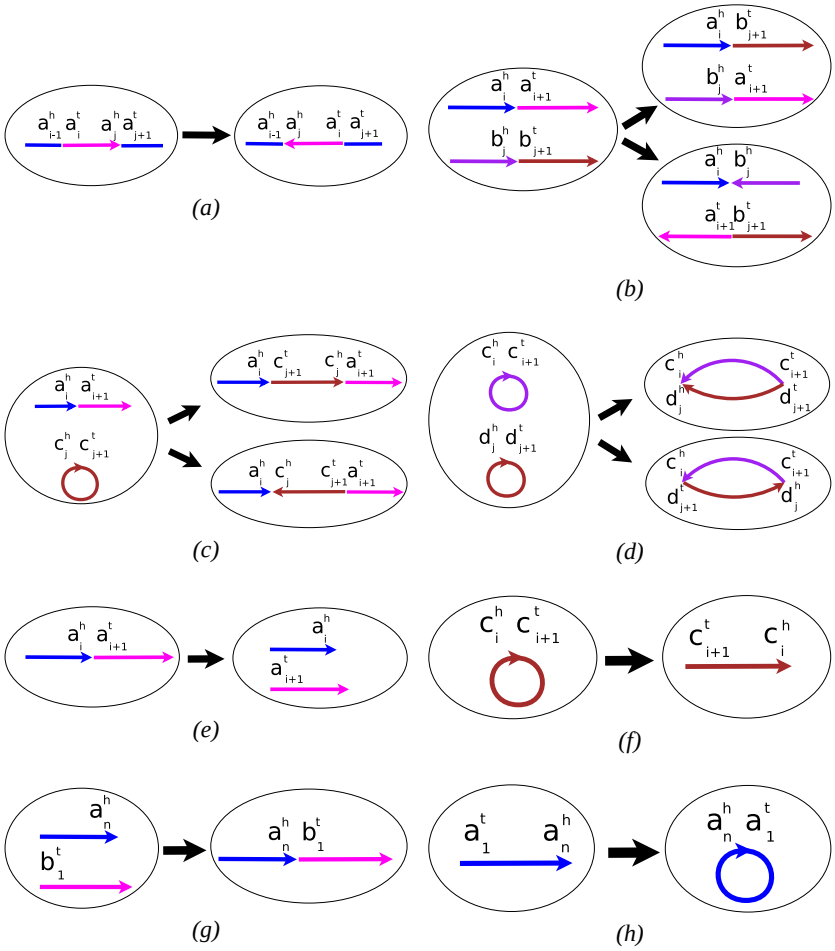


Fig. 1. Possible rearrangements.

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 select two elements uniformly *with replacement* from the multiset of all adjacencies and telomeres and then decide which rearrangement event we apply to these two elements. We have eight cases in all (refer to Fig. 1).

Select two different adjacencies, or one adjacency and one telomere, in the same chromosome (Fig. 1a). 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 $A = (a_1 \dots a_{i-1} a_i \dots a_j a_{j+1} \dots a_n)$. Reversing all genes between a_i and a_j yields $(a_1 \dots a_{i-1}\text{-}a_j \dots \text{-}a_i a_{j+1} \dots a_n)$. 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 $\{a_i^t, a_{j+1}^t\}$. This operation causes an inversion(another possible operation in DCJ model to creat a new circular chromosome is forbidden in our model).

Select two adjacencies, or one adjacency and one telomere, in two linear chromosomes (Fig. 1b). For example, select two adjacencies, $\{a_i^h, a_{i+1}^t\}$ from one linear chromosome $A = (a_1 \dots a_i a_{i+1} \dots a_n)$ and $\{b_j^h, b_{j+1}^t\}$ from another linear chromosome $B = (b_1 \dots b_j b_{j+1} \dots b_m)$. Now exchange the two segments between these two chromosomes C and D. There are two possible outcomes, $(a_1 \dots a_i b_{j+1} \dots b_m)$ and $(b_1 \dots b_j a_{i+1} \dots a_n)$ or $(a_1 \dots a_i\text{-}b_j \dots \text{-}b_1)$ and $(\text{-}b_n \dots \text{-}b_{j+1} a_{i+1} \dots a_n)$. Two adjacencies, $\{a_i^h, a_{i+1}^t\}$ and $\{b_j^h, b_{j+1}^t\}$, are replaced by $\{a_i^h, b_{j+1}^h\}$ and $\{a_{i+1}^t, b_j^t\}$ or $\{a_i^h, b_j^h\}$ and $\{a_{i+1}^t, b_{j+1}^t\}$. This operation causes a translocation.

Select two different adjacencies, or one adjacency and one telomere, in one circular chromosome and one linear chromosome (Fig. 1c). For example, select two adjacencies, $\{a_i^h, a_{i+1}^t\}$ from one linear chromosome $A = (a_1 \dots a_i a_{i+1} \dots a_n)$ and $\{c_j^h, c_{j+1}^t\}$ one circular chromosome $C = (c_1 \dots c_j c_{j+1} \dots c_m)$. Now merge the circular chromosome C into the linear chromosome A. There are two possible outcomes, linear chromosomes $(a_1 \dots a_i c_{j+1} \dots c_m c_1 \dots c_j a_{i+1} \dots a_n)$ or $(a_1 \dots a_i\text{-}c_j \dots \text{-}c_1\text{-}c_m \dots \text{-}c_{j+1} a_{i+1} \dots a_n)$. Two adjacencies, $\{a_i^h, a_{i+1}^t\}$ and $\{c_j^h, c_{j+1}^t\}$, are replaced by $\{a_i^h, c_{j+1}^h\}$ and $\{a_{i+1}^t, c_j^t\}$ or $\{a_i^h, c_j^h\}$ and $\{a_{i+1}^t, c_{j+1}^t\}$. This operation causes a fusion of a circular chromosome with a linear chromosome.

Select two adjacencies in two circular chromosomes (Fig. 1d). For example, select two adjacencies, $\{c_i^h, c_{i+1}^t\}$ from one circular chromosome $C = (c_1 \dots c_i c_{i+1} \dots c_m)$ and $\{d_j^h, d_{j+1}^t\}$ from another circular chromosome $D = (d_1 \dots d_j d_{j+1} \dots d_n)$. Now merge these two circular chromosomes C and D into one new circular chromosome. There are two possible outcomes, circular chromosomes $(c_1 \dots c_i d_{j+1} \dots d_m d_1 \dots d_j c_{i+1} \dots c_m)$ or $(c_1 \dots c_i\text{-}d_j \dots \text{-}d_1\text{-}d_m \dots \text{-}d_{j+1} c_{i+1} \dots c_m)$. Two adjacencies, $\{c_i^h, c_{i+1}^t\}$ and $\{d_j^h, d_{j+1}^t\}$, are replaced by $\{c_i^h, d_{j+1}^h\}$ and $\{c_{i+1}^t, d_j^t\}$ or $\{c_i^h, d_j^h\}$ and $\{c_{i+1}^t, d_{j+1}^t\}$. This operation causes a fusion of two circular chromosomes.

Select the same adjacency twice in one linear chromosome (Fig. 1e). For example, select the adjacency $\{a_i^h, a_{i+1}^t\}$ twice from linear chromosome $A = (a_1 \dots a_i a_{i+1} \dots a_n)$. Then split C into two new linear chromosomes, $(a_1 \dots a_i)$ and $(a_{i+1} \dots a_n)$. The adjacency $\{a_i^h, a_{i+1}^t\}$ is replaced by two telomeres $\{a_i^h\}$ and $\{a_{i+1}^t\}$. This operation causes a fission of a linear chromosome.

Select the same adjacency twice in one circular chromosome (Fig. 1f). For example, select the adjacency $\{c_i^h, c_{i+1}^t\}$ twice from circular chromosome $C = (c_1 \dots c_i c_{i+1} \dots c_m)$. Then linearize C into a linear chromosome, $(c_{i+1} \dots c_m c_1 \dots c_i)$. The adjacency $\{c_i^h, c_{i+1}^t\}$ is replaced by two telomeres $\{c_i^h\}$ and $\{c_{i+1}^t\}$. This operation causes a linearization of a circular chromosome.

Select two telomeres in two linear chromosomes (Fig. 1g). For example, select telomeres $\{a_n^h\}$ and $\{b_1^t\}$ from two different linear chromosomes $A =$

$(a_1 \ldots a_i a_{i+1} \ldots a_n)$ and $B = (b_1 \ldots b_j b_{j+1} \ldots b_m)$. Then concatenate these two linear chromosomes into a single new chromosome $(a_1 \ldots a_i a_{i+1} \ldots a_n b_1 \ldots b_j b_{j+1} \ldots b_m)$. Two telomeres, $\{a_n^h\}$ and $\{b_1^t\}$, are replaced by one adjacency $\{a_n^h, b_1^t\}$. This operation causes a fusion of two linear chromosomes.

Select two telomeres in one linear chromosome (Fig. 1h).[1] For example, select telomeres $\{a_1^t\}$ and $\{a_n^n\}$ from linear chromosome $A = (a_1 \ldots a_i a_{i+1} \ldots a_n)$ (See Fig. 1h). Then circularize the linear chromosome by connecting its two ends. Two telomeres, $\{a_1^t\}$ and $\{a_n^h\}$, are replaced by by one adjacency, $\{a_1^t, a_n^h\}$. This operation causes a circularization of a linear chromosome.

As mentioned earlier, we do not include a fission that creates a circular intermediate. This decision is based on outcomes, not a mechanism; as is the DCJ model itself: that is, the model operations may or may not correspond to actual evolutionary events, but running the model produces simulated genomes that more closely resemble actual genomes.

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 $\{b_j^h, b_{j+1}^t\}$. 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 $\{b_j^h, b_{j+1}^t\}$ is removed, and $L+1$ new adjacencies, $\{b_j^h, a_{i+1}^t\}$, $\{a_{i+1}^h, a_{i+2}^t\}$, …, $\{a_{i+L}^h, b_{j+1}^t\}$, are added. For gene loss, we uniformly select one gene from the set of all candidate genes and delete it, restricting gene loss to the deletion of a single gene copy at a time, following Lynch [15]. 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, $\{a_{i-1}^h, a_i^t\}$ and $\{a_i^h, a_{i+1}^t\}$, are replaced by one adjacency, $\{a_{i-1}^h, a_{i+1}^t\}$.

3 Results

3.1 Model restricted to rearrangements

Edit distance computation

The edit distance between two genomes is the minimum number of allowed evolutionary operations necessary to transform one genome into the other.

Consider two genomes with equal gene content and no duplicate genes. If both genomes consist of only linear chromosomes, the model of Hannenhalli and

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

Pevzner [6] allows the computation of the edit distance under inversions, translocations, fusions, and fissions, hereafter the *HP distance*. The edit distance in our model can be no larger than the HP distance, since our model includes all operations in the HP model and more (circularizations and linearizations).

In fact, the two edit distances are equal for genomes composed of only linear chromosomes. Suppose there are intermediate circular chromosomes in some sorting path in our model; then we can always find pairs of operations, one to circularize a linear chromosome and the other to linearize that circular chromosome, that can be replaced by a fission and a fusion in the HP model. So any optimal sorting path in our model can be transformed into an optimal sorting path of equal length in the HP model; therefore the edit distance in our model is equal to the HP distance—although the number of optimal sorting paths may be different.

If two genomes have both linear and circular chromosomes, the edit distance in our model can be no smaller than the DCJ distance [2], since the DCJ model includes all operations in our model and more. Bergeron *et al.* [3] gave a linear-time algorithm to compute the extra cost of not resorting to "forbidden" DCJ operations to compute the HP distance; their algorithm also applies to our model. Thus for any two genomes with equal gene content and no duplicate genes, the edit distance in our model can be computed in linear time.

3.2 Genome structure prediction

In this section, we prove that our new model respects the distinction between eukaryotic and prokaryotic genomes.

Theorem 1. *Let the ancestral genome have one circular chromosome with n genes. After $O(n)$ rearrangements events, with probability $1-n^{-O(1)}$, the final genome contains a single circular chromosome or a collection of $O(\log n)$ linear chromosomes.*

Proof. We examine the effect of rearrangements on the genome structure. Given the original genome with one circular chromosome, only one of our eight cases can result in a linearization: *select the same adjacency twice* (Fig. 1f). Once we have only linear chromosomes, two cases can directly result in a change in the number of linear or circular chromosomes: *select the same adjacency twice* (Fig. 1e) and *select two telomeres* (Fig. 1h). The probability for selecting the same adjacency twice is $O(1/n)$; that for selecting two telomeres is $O(t^2/n^2)$, where t is the number of telomeres. Every time we select the same adjacency twice, we increase the number of linear chromosomes by 1. Let the indicator variable X_i represent whether or not we select the same adjacency twice at the ith step and write k for the number of evolutionary events. Set $X = \sum_{i=1}^{k} X_i$ and let μ be the expectation of X. The Chernoff bound shows

$$Pr(X > (1+\delta)\mu) < \left(e^{\delta}/(1+\delta)^{1+\delta}\right)^{\mu}$$

In our case, $k = O(n)$, $\mu = O(1)$, $\delta = O(logn)$, so that we get

$$Pr(X > O(\log n)) < n^{-O(1)}$$

Let the indicator variable Y_i represent whether or not we select two telomeres at the ith step. Since $t = 2X$, t is bounded by $O(\log n)$ with probability $1 - n^{-O(1)}$. Thus, with probability $1 - n^{-O(1)}$, we have

$$Pr(Y_i = 1) < O((\log n)^2/n^2),$$

Now set $Y = \sum_{i=1}^{k} Y_i$. We have

$$Pr(Y > 0) \leqslant \sum_{i=1}^{k} Pr(Y_i = 1) < n^{-O(1)}.$$

Overall, then, with probability $1 - n^{-O(1)}$, $X < O(\log n)$ and $Y = 0$, which means that the final genome structure has either a collection of $O(\log n)$ linear chromosomes or a single circular chromosome. □

Thm. 1 tells us that, if the original genomic structure starts from a circular chromosome, most current genomes will contain a single circular chromosome or a collection of linear chromosomes. However, if the initial genome structure was, e.g., a mix of linear and circular chromosomes, would such a structure be stable through evolution? We can characterize all stable structures in our model under some mild conditions.

Theorem 2. *Let the ancestral genome have n genes and assume that there are positive constants c_1 and α such that each chromosome in the ancestral genome has at least $c_1 n^{\alpha}$ genes. Let c_2 be some constant obeying $c_2 > 2c_1$. After $c_2 n^{1-\alpha} \log n$ rearrangements, with probability $1 - O(n^{-\alpha} \log n)$, the final genome contains either a single circular chromosome or a collection of linear chromosomes.*

Proof. In our evolutionary model, consider the case of selecting two adjacencies or one adjacency and one telomere in two different chromosomes. If one of the two chromosomes is circular, a fusion will merge the circular chromosome into the linear chromosome (Fig. 1c). If both chromosomes are circular, a fusion will merge the two chromosomes into a single circular chromosome (Fig. 1d). We use a graph representation, G, for the genome structure, where each circular chromosome is represented by a vertex A_i and all of the linear chromosomes (if any) are represented by a single vertex B . In the evolutionary process, if two adjacencies or one adjacency and one telomere are selected in two different chromosomes, connect the vertices of these two chromosomes. We first ignore circularizations of linear chromosomes (Fig. 1h), then the genome ends up with a single circular

chromosome or a collection of linear chromosomes if and only if the corresponding graph G is connected finally. We therefore bound the probability that the graph G is not connected after $c_2 n^{1-\alpha}\log n$ rearrangements. If G is not connected, there is at least one bipartition of the vertices into S_1 and S_2 in which no edge has an endpoint in each subset. Assume there are g_1 and g_2 genes in S_1 and S_2, respectively; then $min\{g_1, g_2\} \geqslant c_1 n^{\alpha}$ and $g_1 + g_2 = n$. Since there are at most $\frac{1}{c_1} n^{1-\alpha}$ chromosomes, we can write

$$\begin{aligned} Pr(G \text{ is not connected}) &< \binom{\binom{g_1}{2} + \binom{g_2}{2}}{c_2 n^{1-\alpha}\log n} \Big/ \binom{\binom{g_1+g_2}{2}}{c_2 n^{1-\alpha}\log n} \\ &< (1 - c_1 n^{1-\alpha})^{c_2 n^{1-\alpha}\log n} < O(n^{-2\alpha}) \end{aligned}$$

Let indicator variable X_i represent whether or not we select the same adjacency twice at the ith step (Fig. 1e,f) and set $X = \sum_{i=1}^{c_2 n^{1-\alpha}\log n} X_i$. We have

$$\begin{aligned} &Pr(X_i = 1) \leqslant 1/n \\ &Pr(X > 0) \leqslant \sum_{i=1}^{c_2 n^{1-\alpha}\log n} Pr(X_i = 1) = O(n^{-\alpha}\log n). \end{aligned}$$

Now we bound the probability of selecting two telomeres in the same linear chromosome (Fig. 1h), which causes circularization of this chromosome. For each linear chromosome, there are four possible ways of selecting two corresponding telomeres. Since the number of linear chromosomes l is bounded by $\frac{1}{c_1} n^{1-\alpha}$, there are at most $\frac{4}{c_1} n^{1-\alpha}$ ways to circularize one linear chromosome in all $(n+l)^2$ ways of selecting two adjacencies or telomeres. Again, let indicator variable Y_i represent circularization of one linear chromosome at the ith step and set $Y = \sum_{i=1}^{c_2 n^{1-\alpha}\log n} Y_i$. We have

$$\begin{aligned} Pr(Y > 0) &\leqslant \sum_{i=1}^{c_2 n^{1-\alpha}\log n} Pr(Y_i = 1) \\ &\leqslant 4c_2 \log n / c_1 n^{2\alpha} < O(n^{-2\alpha}\log n) \end{aligned}$$

Thus, with probability $1 - O(n^{-\alpha}\log n)$, we have: G is connected, $X = 0$, and $Y = 0$, so that the final genome contains either a single circular chromosome or a collection of linear chromosomes. □

The restriction on the minimum size of chromosomes in the ancestral genomes is very mild, since the parameter α can be arbitrarily small.

Our model also predicts, for genomes composed of a collection of linear chromsomes, convergence to a certain number of chromosomes, which depends on the total number of genes.

Theorem 3. *Assume there are n genes and fewer than* $\frac{1+\sqrt{1+4n}}{2}$ *linear chromosomes in the original genome. The number of linear chromosomes increases during rearrangements, converging to* $\frac{1+\sqrt{1+4n}}{2}$.

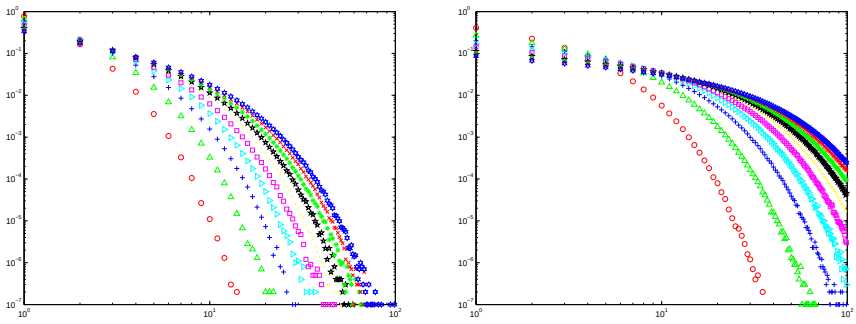
Proof. Assume there are l linear chromosomes in the original genome. In our model, the number of linear chromosomes increases by 1 with probability $\frac{1}{n+l}$ and decreases by 1 with probability $(\frac{l}{n+l})^2$. Since we have $l < \frac{1+\sqrt{1+4n}}{2}$, an increase is more likely. The stable equilibrium follows from the equation $\frac{1}{n+l} = (\frac{l}{n+l})^2$. □

These theorems are not affected by duplications and losses, as long as the latter are reflected in the sizes of chromosomes and the total number of genes.

3.3 Sizes of gene families

Of most concern in a duplication and loss model is the distribution of the sizes of the gene families, since that is one of the few aspects of the process that has been observed to obey general laws. Our sole aim in this section is to demonstrate through simulations that our model, which uses the duplication/loss model of Lynch, yields distributions consistent with what Lynch suggested [15].

Our experiments start with a genome with no duplicated genes. This genome is then subjected to a prescribed number k, varying from from 0 to 10 times the number of genes, of evolutionary events chosen according to p_d and p_l to obtain different genomes G^k. We test a large number of different choices of parameters on varying sizes of genomes; as the results are consistent throughout, we report two cases: (a) 1'000 genes with $L = 10$, $p_d = 0.2$, and $p_l = 0.8$; and (b) 10'000 genes with $L = 10$, $p_d = 0.4$, and $p_l = 0.6$. The data in Fig. 2 summarizes 1'000 runs for each parameter setting. (The parameters chosen correspond to those used



(a) 1'000 genes *(b) 10'000 genes*

Fig. 2. Probability distribution of the size of gene families, for various numbers of events, increasing from the leftmost (*#events* = *#genes*) to the rightmost (*#events* = $10 \times$ *#genes*).

in our distance estimation model [13].) The shape of the distributions of gene family sizes is generally similar to the observations presented by Lynch [15].

4 Discussion and Conclusions

Thm. 1 and Thm. 2 together show that our model respects the distinction between the organization of most prokaryotic genomes (one circular chromosome) and that of most eukaryotic genomes (multiple linear chromosomes). In contrast, the HP model [6] deals with only linear chromosomes, while the DCJ model [2, 22] (assuming uniform distribution of all possible DCJ operations) predicts that over half of modern genomes consisting of only circular chromosomes will have more than one circular chromosome.

There is evidence about the linearization of circular chromosomes during bacterial evolution [20] and the increase in the number of chromosomes of eukaryotic groups by centric fission [8, 9], both of which accord with Thm. 3. According to the minimum interaction theory of Imai *et al.* [10], genome evolution in eukaryotes proceeds as a whole toward increasing the number of chromosomes. Their theory predicts that the highest number of chromosomes in mammals should be 166, while their simulations yield a range of 133–138 for this number [11]. The latter range agrees with our model (as well as the models in [6, 2, 22], if we assume that the two cuts are uniformly selected) if the number of genes is around 20'000, a fairly typical value for mammals.

Fig. 2 shows that our model of gene duplications and losses readily generates distributional forms close to the observations presented by Lynch [15]. Different parameters for gene duplications and losses, and the number of evolutionary events, influence the the distributions of gene family sizes: such information can help us improve the estimation of the actual number of evolutionary events as well as infer the parameters for duplications and losses in our model [13].

According to our model, more rearrangements, gene duplications, and gene losses will linearize circular chromosomes, increase the number of linear chromosomes, and increase the number of genes—i.e., will favor a shift from a prokaryotic architecture to a eukaryotic one. However prokaryotic architectures exist in large numbers today. The reason is to be found in population sizes. In a large population, as with most prokaryotic organisms, most alleles are likely to be eliminated by purifying selection, whereas, in a small population, neutral or even deleterious mutations can be fixated more easily. Thus many forms of mutant alleles that are able to drift to fixation in multicellular eukaryotes are eliminated by purifying selection in prokaryotes as population sizes decreased dramatically in the transition from prokaryotes to multicellular eukaryotes [15]. Similarly, the fixation of rearrangements, gene duplications, and gene losses (all "rare genomic events" [17]) in prokaryotic species is also more difficult compared to that in eukaryotes. Thus, in our model, prokaryotes tend to have one circular chromosome

and a small number of genes, while eukaryotes tend to have multiple linear chromosomes and a large number of genes, in response to a reduction in purifying selection. Our model of gene rearrangement, duplication, and loss is the first to give rise naturally to such a structure; and it does so independently of the choice of parameters, which influence only the tapering rate of the size of gene families.

References

1. D.A. Bader, B.M.E. Moret, & M. Yan. A fast linear-time algorithm for inversion distance with an experimental comparison. *J. Comp. Biol.*, 8(5):483–491, 2001.
2. A. Bergeron, J. Mixtacki, & J. Stoye. A unifying view of genome rearrangements. In *Proc. 6th Workshop Algs. in Bioinformatics (WABI'06)*, number 4175 in Lecture Notes in Comp. Sci., pages 163–173. Springer Verlag, Berlin, 2006.
3. A. Bergeron, J. Mixtacki, & J. Stoye. A new linear-time algorithm to compute the genomic distance via the double-cut-and-join distance. *Theor. Comp. Sci.*, 410(51):5300–5316, 2009.
4. G. Fertin, et al. *Combinatorics of Genome Rearrangements*. MIT Press, 2009.
5. M.W. Hahn, et al. Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Research*, 15(8):1153–1160, 2005.
6. S. Hannenhalli & P.A. Pevzner. Transforming mice into men (polynomial algorithm for genomic distance problems). In *Proc. 36th IEEE Symp. Foundations of Comp. Sci. (FOCS'95)*, pages 581–592. IEEE Press, Piscataway, NJ, 1995.
7. M.A. Huynen & E. van Nimwegen. The frequency distribution of gene family sizes in complete genomes. *Mol. Biol. Evol.*, 15(5):583–589, 1998.
8. H.T. Imai. On the origin of telocentric chromosomes in mammals. *J. Theor. Biol.*, 71(4):619–637, 1978.
9. H.T. Imai & R.H. Crozier. Quantitative analysis of directionality in mammalian karyotype evolution. *American Naturalist*, 116(4):537–569, 1980.
10. H.T. Imai, et al. Theoretical bases for karyotype evolution. 1. the minimum-interaction hypothesis. *American Naturalist*, 128(6):900–920, 1986.
11. H.T. Imai, et al. Estimation of the highest chromosome number of eukaryotes based on the minimum interaction theory. *J. Theor. Biol.*, 217(1):61–74, 2002.
12. E.V. Koonin, Y.I. Wolf, & G.P. Karev. The structure of the protein universe and genome evolution. *Nature*, 420:218–223, 2002.
13. Y. Lin, V. Rajan, K.M. Swenson, & B.M.E. Moret. Estimating true evolutionary distances under rearrangements, duplications, and losses. In *Proc. 8th Asia-Pacific Bioinf. Conf. APBC'10*, volume 11 (Suppl. 1), S54, of *BMC Bioinformatics*, 2010.
14. N.M. Luscombe, et al. The dominance of the population by a selected few: power-law behaviour applies to a wide variety of genomic properties. *Genome Biology*, 3(8), 2002.
15. M. Lynch. *The Origins of Genome Architecture*. Sinauer, 2007.
16. J. Qian, N.M. Luscombe, & M. Gerstein. Protein family and fold occurrence in genomes: power-law behavior and evolutionary model. *J. Mol. Biol.*, 313:673–681, 2001.
17. A. Rokas & P.W.H. Holland. Rare genomic changes as a tool for phylogenetics. *Trends in Ecol. & Evol.*, 15:454–459, 2000.
18. A.H. Sturtevant. A case of rearrangement of genes in Drosophila. *Proc. Nat'l Acad. Sci., USA*, 7:235–237, 1921.
19. A.H. Sturtevant & T. Dobzhansky. Inversions in the third chromosome of wild races of Drosophila pseudoobscura, and their use in the study of the history of the species. *Proc. Nat'l Acad. Sci., USA*, 22:448–450, 1936.
20. J.-N. Volff & J. Altenbuchner. A new beginning with new ends: Linearisation of circular chromosomes during bacterial evolution. *FEMS Microbiol. Lett.*, 186:143–150, 2000.
21. I. Yanai, C.J. Camacho, & C. DeLisi. Predictions of gene family distributions in microbial genomes: evolution by gene duplication and modification. *Phys. Rev. Lett.*, 85(12):2641–2644, 2000.
22. S. Yancopoulos, O. Attie, & R. Friedberg. Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics*, 21(16):3340–3346, 2005.