Structural vs. functional mechanisms of duplicate gene loss following whole genome doubling

David Sankoff, Baoyong Wang, Chunfang Zheng
Department of Mathematics and Statistics
University of Ottawa
585 King Edward Avenue,
Ottawa, ON, Canada, K1N 6N5
Email: {sankoff,bwang077,czhen033}@uottawa.ca

Carlos Fernando Buen Abad Najar
Facultad de Ciencias
Universidad Nacional Autónoma de México
Avenida Universidad 3000
Distrito Federal, México
Email: fernandobuenabad@ciencias.unam.mx

The process of whole genome doubling (WGD) gives rise to two copies of each chromosome in a genome, containing the same genes in the same order. Through an attrition mechanism known as fractionation, one of each pair of duplicate genes is lost over evolutionary time, resulting in an interleaving patterns of deletions from duplicated regions [1]. This differentiates the WGD/fractionation model from general approaches to gene duplication, pioneered by El-Mabrouk [2].

An important biological controversy in evolutionary theory arises of whether duplicated genes are deleted through random excision – elimination of excess DNA – namely the deletion of chromosomal segments containing one or more genes [3], which we term the "structural" mechanism, or through gene-by gene events such as epigenetic silencing and pseudogenization [4], which are "functional" mechanisms.

This debate may be formulated in terms of deletion events removing a number X of contiguous genes, where X is drawn from a geometric distribution γ with mean μ. Here the one-at-a-time deletion model is represented by $\mu = 1$, while the random number of deletions at a time holds if $\mu > 1$.

In this paper, we investigate the discrimination problem of choosing between the two models based on deletion run-length statistics (resulting from overlapping deletion events). This involves comparing an observed genome containing single-copy genes, originally members of duplicate pairs, to the predictions of the models for $\mu = 1$ and for $\mu > 1$. This requires knowledge of the run-length distribution, given a total number of deleted genes and remaining duplicate pairs. While this is easily calculated for the case $\mu = 1$, the the distribution for the opposing scenario $\mu > 1$ is not known.

For modeling purposes, we consider a doubled genome made up of a pair of identical linear chromosomes each containing genes $g_1, \ldots, g_N$. At each time $t = 1, 2, \ldots$, one such doubled gene g_i is chosen at random, and a value a is chosen from a geometric distribution γ with mean μ. Then $g_i, g_{i+1}, \ldots, g_{i+a-1}$ are deleted from one of the genomes – they become single-copy genes – unless some of these are already single-copy. In the latter case, we skip existing single-copy genes and proceed to convert the next double-copy genes we encounter until a total of a double-copy genes have been converted to single-copy. This model is biologically realistic, although for simplicity, we assume all deletions take place from one and the same genome. In a more complete model, deletion events occur on one or the other chromosome, with probabilities ϕ and $1 - \phi$ [5].

Overlapping deletion events and skipping result in the creation of runs of single-copy genes whose length is the sum of a number of geometric variables, which are not, however, i.i.d., and thus do not produce a negative binomial distribution. The run lengths of the remaining double-copy genes is geometrically distributed with a probability distribution ρ_t, with a mean ν_t that decreases with t [5], [6].

An attempt to determine ψ_t analytically starts with the calculation of how many deletion events have overlapped to form a run of single-copy genes at time t. We can derive a formula to predict whether a deletion event would create a new run of single-copy genes, probability p_0; overlap exactly one existing run, thus extending it without changing the total number of runs, probability p_1; overlap two runs, producing one larger combined run in place of the two pre-existing ones, probability p_2; and so on. These probabilities all depend solely on γ and ρ_t. For example, we examine the case of p_0.

The proportion of terms in runs of length l is $l\rho_t(l)/\nu_t$, where $\nu_t = \sum_{l>0} l\rho_t(l)$. The probability p_0 that a deletion event falls within a run of double-copy genes without deleting the terms at either end is

$$\begin{aligned} p_0 &= \sum_{l>2} \frac{l\rho_t(l)}{\nu_t} \sum_{j=2}^{l-1} \frac{1}{l} \sum_{a=1}^{l-j} \gamma(a) \\ &= \frac{1}{\nu_t} \sum_{l>2} \rho_t(l) \sum_{a=1}^{l-2} (l-a-1)\gamma(a) \end{aligned} \quad (1)$$

where j indexes the starting position of the deletion within a run of length l, and a is the number of terms deleted.

This formula requires quadratic computing time, but the p_i for higher i, require polynomial time of degree $i+2$. These probabilities, however, can in fact be reduced to closed form, so that computing time is a negligible constant. In lieu of the detailed calculations, here present the continuous version of the deletion process, which is simpler. In this case, the two identical chromosomes at time $t = 0$ are linear segments. At each time $t = 1, 2, \ldots$, a random point g is chosen on the chromosome, and a value X is chosen from an exponential distribution $f(a) = \frac{1}{\mu} e^{-\frac{a}{\mu}}, \quad a \geq 0$, with mean μ. If $X = a$, then the segment $[g, g+a]$ is deleted from one of the genomes

978-1-4673-1321-6/12/$31.00

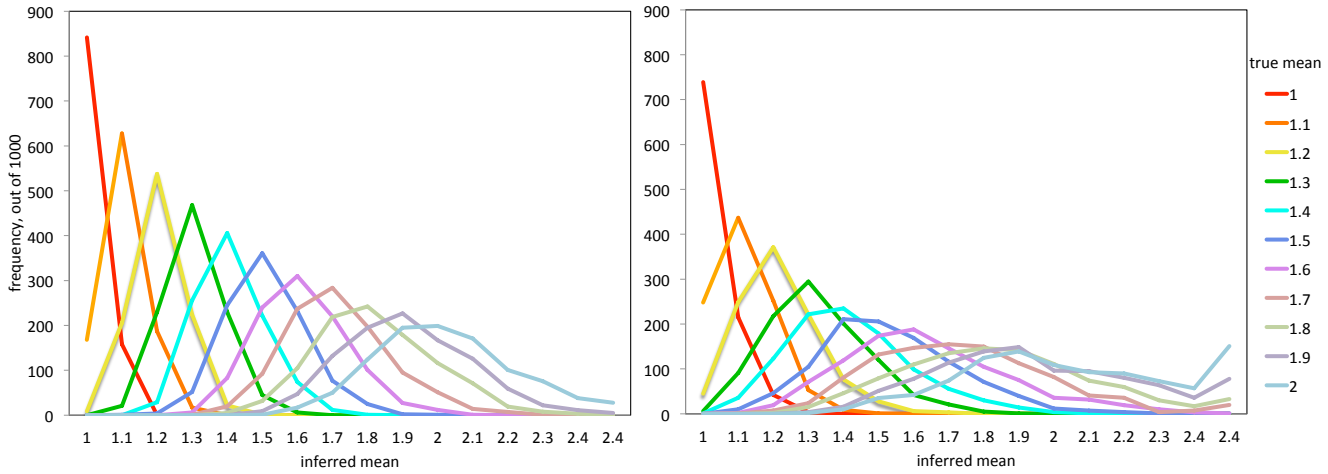

Fig. 1. Frequency of $\hat{\mu}$, the value for which $D^{(S_i)}_{\mu,N,1-\theta}$ between the sample cumulative and the distribution $F_{\mu,N,1-\theta}$ is minimal. All data involve a proportion of $1-\theta = 0.20$ deleted genes. Left: $N = 900$, right: $N = 300$.

– $[g, g+a]$ becomes a single-copy region – unless part of it is already single-copy. In the latter case, we skip existing single-copy regions and proceed to convert the next double-copy region we encounter until a total measure a of double-copy regions have been converted to single-copy.

In analogy with ρ_t in the discrete model, the lengths of the remaining double-copy segments follow an exponential distribution σ_t, with a mean ν_t that decreases with t.

The proportion of undeleted regions accounted for by segments of length ldl is $\frac{l\sigma(l)}{\nu_t}dl$, where $\nu_t = \int_0^\infty l\sigma(l)dl$. Then the probability p_0 that a deletion event falls completely within an undeleted segment is

$$p_0 = \int_{l=0}^{\infty} \frac{l\sigma_t(l)}{\nu_t} \int_{x=0}^{l} \frac{1}{l} \int_{y=0}^{l-x} f(y)dy\,dx\,dl \qquad (2)$$

$$= \frac{\nu_t}{\mu+\nu_t}. \qquad (3)$$

We can prove by induction that the probability a deletion event overlaps exactly q existing runs of deletions is:

$$p_q = \frac{\nu_t}{\mu+\nu_t}\left(\frac{\mu}{\mu+\nu_t}\right)^q. \qquad (4)$$

Thus we have the surprisingly uncomplicated result that the number q of pre-existing runs of single-copy regions overlapped by a new deletion event is geometrically distributed on $q = 0, 1, \ldots$ with parameter $\mu/(\mu+\nu_t)$.

Although having a closed form for p_q constitutes progress towards the computation of the run-length distribution ψ_t, or eventually towards some analytical results on it, how to find this distribution remains a difficult question. In the interim, we may use simulations to study the discrimination problem.

For various combinations of the parameters μ, N and $1-\theta$, we first estimated quite precisely (through 1000 simulations) the cumulative distribution $F_{\mu,N,1-\theta}$ of run lengths for single-copy regions. Once these cumulative distributions were established, we then carried out a discrimination study. For each value of μ and N, we sampled 1000 new individual trajectories of the deletion process at various values of $1-\theta$. For each value of $1-\theta$, we set up "bins" corresponding to the fifteen values of μ for which we had constructed cumulatives. Then for each sample S_i, we constructed the cumulative distribution of runs of deleted genes of length $1, 2, \ldots$. We calculated the Kolmogorov–Smirnov statistic $D^{(S_i)}_{\mu,N,1-\theta}$ between the sample cumulative and the distribution $F_{\mu,N,1-\theta}$ for each fifteen values of μ and assigned the sample to the bin corresponding to the minimal value of D, which we called $\hat{\mu}$ for that sample.

Figure 1 shows the distributions of $\hat{\mu}$ for the 1000 samples $S_1, \ldots, S_{1000}$, for $N = 900$ and $N = 300$. A separate distribution is drawn for each of the trial values of μ used to generate the samples. The complete set of these distributions can be used to to address the original problem of discriminating between the gene-by-gene "functional model" ($\mu = 1$) and the random excision "structural" model ($\mu > 1$). The distributions are more dispersed for smaller N, and for higher values of μ and $1-\theta$.

Acknowledgment

Research supported in part by grants from the Natural Sciences and Engineering Research Council of Canada. DS holds the Canada Research Chair in Mathematical Genomics.

References

[1] Wolfe KH, Shields DC: Molecular evidence for an ancient duplication of the entire yeast genome. Nature 1997, 387:708–13.

[2] El-Mabrouk, N: Genome rearrangement by reversals and insertions/deletions of contiguous segments. in Combinatorial Pattern Matching, 11th Annual Symposium, 2000. Lecture Notes in Computer Science 1848:222–234.

[3] van Hoek MJ, Hogeweg P: The role of mutational dynamics in genome shrinkage. Molecular Biology and Evolution 2007, 24:2485–2494.

[4] Byrnes JK, Morris GP, Li WH: Reorganization of adjacent gene relationships in yeast genomes by whole-genome duplication and gene deletion. Molecular Biology and Evolution 2006, 23:1136–1143.

[5] Sankoff D, Zheng C, Wang B: A model for biased fractionation after whole genome duplication. BMC Genomics 2012, 13:S1, S8.

[6] Wang B, Zheng C, Sankoff D: Fractionation statistics. BMC Bioinformatics 2011, 12: S9, S5.