B Appendix: Crossover Behavior

For nonzero R and times other than zero and infinity, the fold distribution will not be strictly exponential, nor will it conform to the limiting distribution (8). For small times, we would intuitively expect the histogram to be dominated by duplication events involving the initial N_0 genes. This is confirmed by the behavior of the analytic solution for small t:

$$\begin{aligned} F(m,t) \approx & N_0 \left(1 - \frac{t}{N_0}\right) \left(\frac{t}{N_0}\right)^{m-1} \\ & + A_m \left[1 + \frac{R+1}{N_0} t - \left(1 - \frac{t}{N_0}\right)^m\right] - \sum_{i=1}^{m-1} A_i \beta_{m-i}^i \left(1 - \frac{t}{N_0}\right)^i \left(\frac{t}{N_0}\right)^{m-i} \end{aligned} \tag{32}$$

From this approximation, it is clear that the terms involving N_0 dominate for small times. Consequently, the fold distribution will resemble an exponential distribution more than the limiting distribution early on in the evolution of the genome. It is also clear that the histogram $F(m,t)$ will not approach the limiting distribution uniformly; the rate of convergence will depend on cluster size.

There are many possible ways of characterizing this transformation of the fold distribution, each suggesting a different notion of a "'crossover" time. We have looked at the convergence of the probability distribution as a whole. To quantify the extent to to which the actual distribution $p(m)$ resembles a second distribution, say $p_A(m)$, we adopt the sum of the squared differences as our metric:

$$\eta_A = \sum_m (p(m) - p_A(m))^2 \tag{33}$$

$$\sum_m p(m) = \sum_m p_A(m) = 1 \tag{34}$$

Figure 7 tracks the evolution of $p(m)$ according to this metric when $R = 1.0$ and $N_0 = 100$. At each time, the closeness of $p(m)$ to the limiting distribution (8) is shown, as is the closeness to the best fitting exponential distribution for that time, obtained by a least-squares regression of $logp$ against m. For times greater than $t \approx 70$, the distribution of fold sizes resembles the final distribution more than any exponential distribution, this defines the crossover time for this set of parameters. The sum extends to cluster sizes large enough to ensure numerical convergence.

Figure 8 plots the crossover time as a function of R for two values of N_0. The range of R is chosen so that new fold acquisitions occur less frequently than (or as often as) gene duplication. The crossover time displays two distinct regimes. Within each regime it is approximately inversely proportional to R and directly proportional N_0. A different proportionality constant applies in each regime: $T_c \sim N_0/R$. These numerical results confirm that crossover occurs roughly when the number of new fold introductions: RT_c becomes comparable to the initial genome size N_0. The details of the dependence are not that important, as they are no doubt strongly affected by the choice of metric.

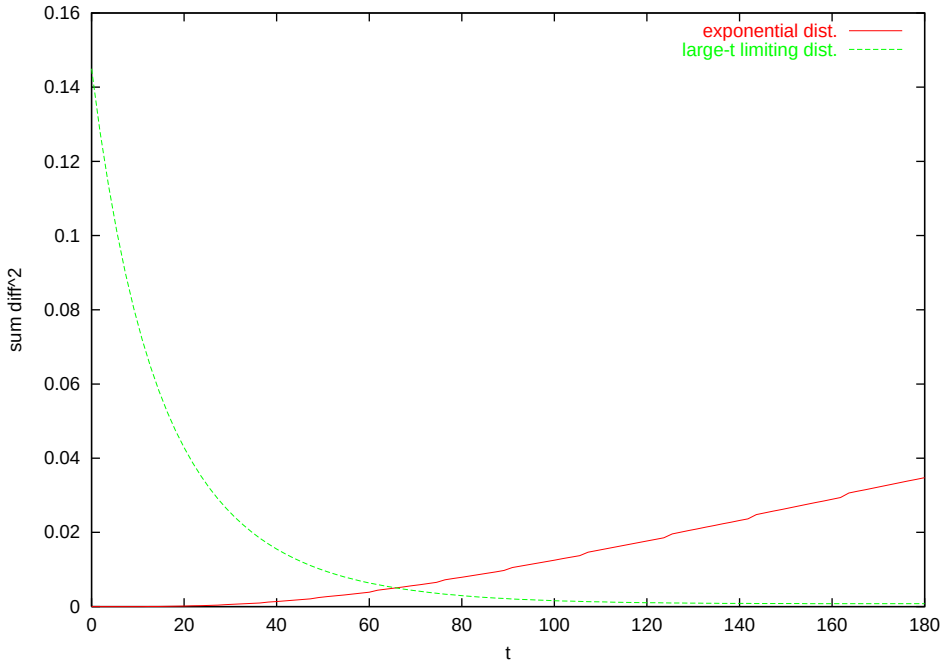
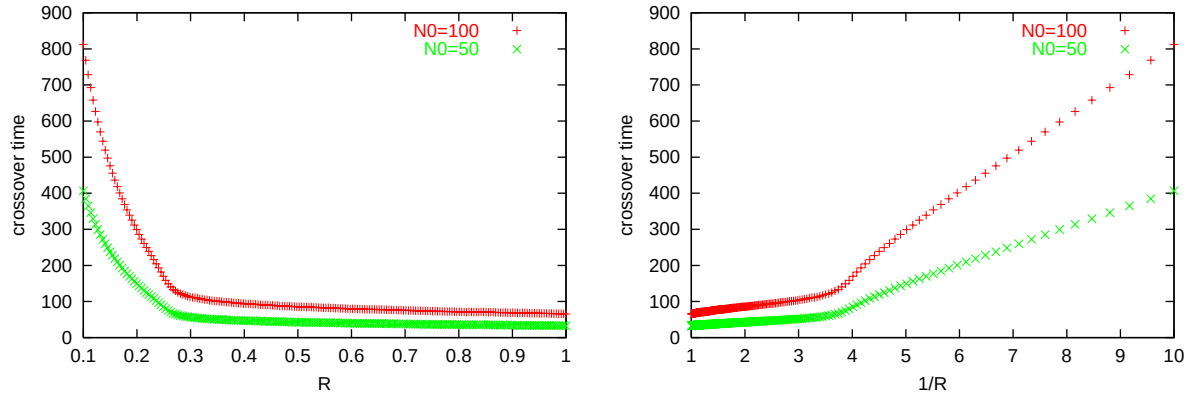


Figure 7: Crossover from exponential to large-time limiting distribution for $R = 1.0$ and $N_0 = 100$.



(a) (b)

Figure 8: Crossover time for $N_0 = 100$ and $N_0 = 50$, plotted as a function of (a) R, and (b) $1/R$.