G Appendix: The Effects of Selection Pressure

Recall that we have assumed that there are only two duplication types: type "A" and type "B", and that "B" genes are γ times more likely to be chosen for duplication than "A" genes. There will still be one duplication event, on average, per unit time, so the total expected number of genes will remain the same, but the allocation of the total between types "B" and "A" will depend on γ. We will assume that $\gamma > 1$, so it is the "B" types that are more likely to be duplicated.

To keep track of the fold population we now need two histograms: $F_A(m,t)$ and $F_B(m,t)$ to distinguish between the duplication types. The full fold histogram is the sum of both sub-histograms: $F(m,t) = F_A(m,t) + F_B(m,t)$. Similarly, let $G_A(t)$ and $G_B(t)$ represent the total number of genes for each type and define a new variable $G_\gamma(t)$:

$$G_\gamma(t) = G_A(t) + \gamma G_B(t) \tag{67}$$

The evolution equations that extend (2) are:

$$\begin{aligned}
\frac{\partial F_A(m,t)}{\partial t} &= \frac{(m-1)F_A(m-1,t)}{G_\gamma(t)} - \frac{mF_A(m,t)}{G_\gamma(t)} \quad (m>1)\\
\frac{\partial F_A(1,t)}{\partial t} &= R_A - \frac{F_A(1,t)}{G_\gamma(t)}\\
\frac{\partial F_B(m,t)}{\partial t} &= \gamma\frac{(m-1)F_B(m-1,t)}{G_\gamma(t)} - \gamma\frac{mF_B(m,t)}{G_\gamma(t)} \quad (m>1)\\
\frac{\partial F_B(1,t)}{\partial t} &= R_B - \gamma\frac{F_B(1,t)}{G_\gamma(t)}
\end{aligned} \tag{68}$$

Note that we allow new folds to be acquired at different rates for each type: R_A can be different from R_B although we will restrict our numerical examples to the when they are equal.

As before, we derive equations for the total number of genes from the full dynamics (68):

$$\begin{aligned}
\frac{\partial G_A(t)}{\partial t} &= \frac{\partial}{\partial t}\sum_{m=1} mF_A(m,t) = R_A + \frac{G_A(t)}{G_A(t)+\gamma G_B(t)}\\
\frac{\partial G_B(t)}{\partial t} &= \frac{\partial}{\partial t}\sum_{m=1} mF_B(m,t) = R_B + \gamma\frac{G_B(t)}{G_A(t)+\gamma G_B(t)}\\
\frac{\partial G(t)}{\partial t} &= \frac{\partial G_A(t)}{\partial t} + \frac{\partial G_B(t)}{\partial t} = R_A + R_B + 1
\end{aligned} \tag{69}$$

This confirms that the overall duplication rate is still one gene per unit time. The evolution of $G_\gamma(t)$ is more complicated:

$$\frac{\partial G_\gamma(t)}{\partial t} = R_A + \gamma R_B + 1 + \gamma\left[1 - \frac{G(t)}{G_\gamma(t)}\right] \tag{70}$$

It is possible to establish the distributional properties of the genome without having to solve (70) explicitly for the special parameter values encountered previously: (1) the case when there is no introduction of new folds, so $R_A = R_B = 0$; and (2) the limiting distribution when $t \to \infty$. When there is no introduction of new folds, a simple extension of the repeated integration employed in Appendix A establishes that the each of the sub-histograms $F_A(m,t)$ and $F_B(m,t)$ follows an exponential distribution for all times:

$$\begin{aligned}
F_A(m,t) &= N_0^A \exp(-u(t))\left[1-\exp(-u(t))\right]^{m-1}\\
F_B(m,t) &= N_0^B \exp(-\gamma u(t))\left[1-\exp(-\gamma u(t))\right]^{m-1}
\end{aligned} \tag{71}$$

The number of distinct folds of each type, present at $t = 0$ is given by N_0^A and N_0^B. The variable $u(t)$ is determined by $G_\gamma(t)$:

$$u(t) = \int_0^t \frac{ds}{G_\gamma(s)} \tag{72}$$

The full histogram is consequently a sum of exponential distributions:

$$\begin{aligned} p(m,t) &= \frac{F_A(m,t) + F_B(m,t)}{\sum_i F_A(i,t) + F_B(i,t)} \\ &= \frac{N_0^A}{N_0^A + N_0^B} e^{-u} \left[1 - e^{-u}\right]^{m-1} + \frac{N_0^B}{N_0^A + N_0^B} e^{-\gamma u} \left[1 - e^{-\gamma u}\right]^{m-1} \end{aligned} \tag{73}$$

The the large time behavior of the solution is much easier to derive than an exact solution. For large t, $G_\gamma(t)$ will grow linearly with time: $G_\gamma \sim C_\gamma t$, according to a constant C_γ that depends on the rate of fold acquisition and the differential rate of duplication:

$$C_\gamma = \frac{1}{2}\left(R_A + \gamma R_B + 1 + \gamma\right) + \frac{1}{2}\sqrt{(R_A + \gamma R_B + 1 + \gamma)^2 - 4\gamma(R_A + R_B + 1)} \tag{74}$$

In a similar fashion, we define coefficients C_m^A and C_m^B, akin to the coefficients A_m of the solution to the minimal model (30), that describe the ultimate linear growth of the histogram bins: $F_A(m,t) \sim C_m^A t$, and similarly for $F_B(m,t)$. The form of the coefficients is very similar to the minimal model's A_m:

$$\begin{aligned} C_m^A &= \frac{R_A}{C_\gamma + 1} \prod_{i=1}^{m-1} \frac{i}{C_\gamma + i + 1} \\ C_m^B &= \frac{R_B}{C_\gamma + \gamma} \prod_{i=1}^{m-1} \frac{i\gamma}{C_\gamma + \gamma(i+1)} \end{aligned} \tag{75}$$

The normalized probability distribution corresponding to this limit can be found using the same normalization identity that was helpful in deriving the probability distribution in the minimal model (Appendix H):

$$\begin{aligned} p(m,t) &= \frac{C_m^A + C_m^B}{\sum_i C_i^A + C_i^B} \\ &= \frac{C_\gamma}{C_\gamma + 1} \frac{R_A}{R_A + R_B} \prod_{i=1}^{m-1} \frac{i}{C_\gamma + i + 1} + \frac{C_\gamma}{C_\gamma + \gamma} \frac{R_B}{R_A + R_B} \prod_{i=1}^{m-1} \frac{i\gamma}{C_\gamma + \gamma(i+1)} \end{aligned} \tag{76}$$

We have also briefly considered the case of more than two duplication types. When there is no introduction of new folds into the genome, the same argument behind equations (71) and (16) generalizes: the sub-histogram for each duplication type is exponential. Furthermore, we have confirmed numerically that the terminal distribution is not dramatically affected by selection pressure, even when there are several families with significantly different rates of duplication. One particular example, involving a four duplication types appears in Figure (10). In this rather extreme case, types “B”, “C” and “D” are 4.0, 8.0 and 16.0 times more likely to be duplicated than type “A”. The total rate of new fold acquisition is the same for both genomes.

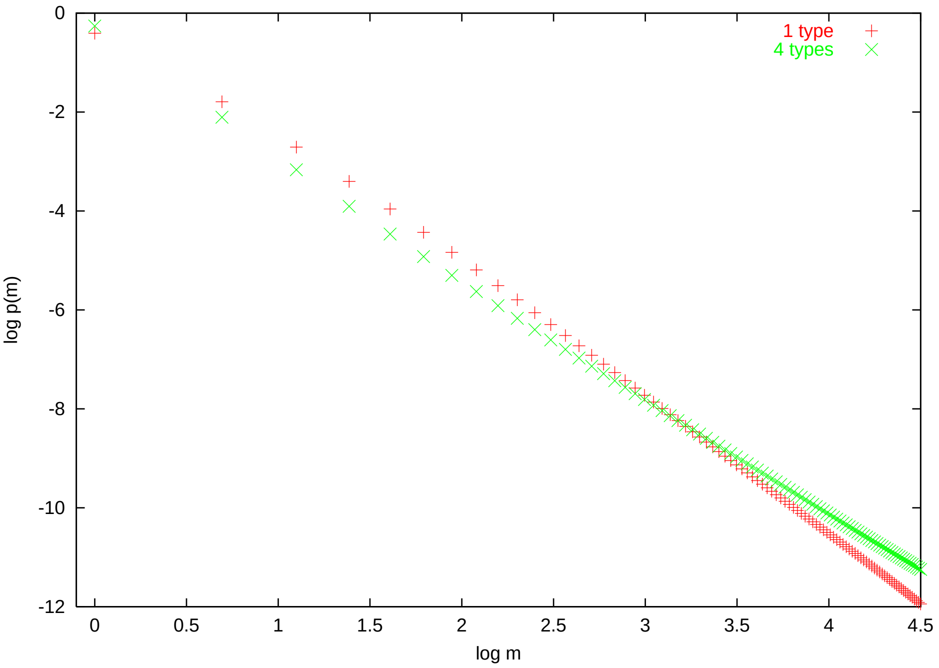

Figure 10: Large time limit for the fold probability distribution for the minimal model (one duplication type) and four duplication types: $\gamma_B = 4, \gamma_C = 8, \gamma_D = 16$. The total rate of new fold acquisition is the same for both genomes.