F Appendix: Perturbation Theory Approximation for the Extended Model

As before, relate time and the number of genes through $\phi(t)$:

$$\phi(t) = 1 + \frac{(R+1-Q)t}{N_0}, \tag{49}$$

This extends the previous definition (5); the variable u keep is still defined as before (20): $u = \log \phi(t)$.

Recall that when $Q = 0$ and $R > 0$ the long-term behavior of $F(m,t)$ is determined by the coefficients A_m, as shown in equation (6). Assume that the large-time solution in the presence of gene deletion is determined by new coefficients B_m:

$$F(m,t) \to B_m \phi(t) = B_m \exp(u) \text{ as } t \to \infty \tag{50}$$

Substituting this ansatz into the fundamental equations (10) leads to:

$$\begin{aligned}(1+R-Q)B_1 &= RN_0 - (1+Q)B_1 + 2QB_2 \\ (1+R-Q)B_m &= (m-1)B_{m-1} - (1+Q)mB_m + Q(m+1)B_{m+1}\end{aligned} \tag{51}$$

Motivated by the numerical results, we will develop the perturbation around a new variable γ_m:

$$B_m = \gamma_m A_m \tag{52}$$

that relates B_m to the $Q = 0$ solution (A_m) as closely as possible. Using the explicit form of A_m from (30) in (51) leads to:

$$\begin{aligned}\gamma_1 &= 1 + \frac{2}{(R+2)(R+3)}\gamma_2 \\ \gamma_m &= \gamma_{m-1} + Q\frac{(1-m)}{R+1+m}\gamma_m + Q\frac{m(m+1)}{(R+1+m)(R+2+m)}\gamma_{m+1}\end{aligned} \tag{53}$$

It is easy to see that when $Q = 0$, $\gamma_m = 1$, which means $B_m = A_m$ for all m.

The perturbation theory approach expands γ_m for each m as a power series in Q:

$$\gamma_m = \sum_{i=0}^{\infty} Q^i \gamma_m^{(i)} \tag{54}$$

From the solution when $Q = 0$ we immediately know the first term in the expansion: $\gamma_m^{(0)} = 1$. The remaining terms are determined order-by-order by substituting into (53) and collecting terms with the same power of Q:

$$\begin{aligned}\gamma_1^{(i)} &= 1 + \frac{2}{(R+2)(R+3)}\gamma_2^{(i-1)} \\ \gamma_m^{(i)} &= \gamma_{m-1}^{(i)} + \frac{(1-m)}{R+1+m}\gamma_m^{(i-1)} + \frac{m(m+1)}{(R+1+m)(R+2+m)}\gamma_{m+1}^{(i-1)}\end{aligned} \tag{55}$$

The first-order ($i = 1$) equations are easy to solve since the the zeroth-order solutions are just unity:

$$\begin{aligned}\gamma_1^{(1)} &= 1 + \frac{2}{(R+2)(R+3)} \\ \gamma_m^{(1)} &= \gamma_1^{(1)} + \sum_{i=2}^{m} g(i) \\ g(i) &= \frac{2+2R+R^2}{1+R+i} - \frac{2+3R+R^2}{2+R+i}\end{aligned} \tag{56}$$

(The function $g(i)$ comes from simplifying the addition of the fractions multiplying $\gamma_m^{(i-1)}$ and $\gamma_{m+1}^{(i-1)}$ in (56)).

An important limitation of the perturbation expansion is revealed by the first order solution. Consider the behavior of the sum:

$$\begin{aligned}\sum_{i=2}^{m} g(i) &\approx \int_2^m dx g(x) \\ &= (2+2R+R^2)\log\frac{m+R+1}{R+3} - (2+3R+R^2)\log\frac{m+R+2}{R+4}\end{aligned} \tag{57}$$

For large m, the sum increases in magnitude logarithmically with m:

$$\sum_{i=2}^{m} g(i) \approx -R\log m \tag{58}$$

This means that no matter how small Q is, for large enough m the first order expansion will fail. This reflects a limitation of the perturbation expansion itself for this problem – stopping the expansion at any finite order will lead to a series valid only up to some maximum size m.

The only way to obtain a consistent expansion is to sum all orders of the series. Unfortunately, the equations (55) are difficult to solve exactly, and even if they were possible to solve, it would be even more difficult to carry out the summation. However, it isn't difficult to figure out the dominant contribution at each order. It helps to first look at the equations for $i = 2$:

$$\begin{aligned}\gamma_m^{(2)} &= \gamma_{m-1}^{(2)} + g(m)\sum_{i=2}^{m} g(i) + \frac{m(m+1)}{(R+1+m)(R+2+m)}g(m+1) \\ \Longrightarrow \gamma_m^{(2)} &= \gamma_1^{(2)} + \sum_{i=2}^{m} g(i)\sum_{j=2}^{i} g(j) + \sum_{i=2}^{m}\frac{i(i+1)}{(R+1+i)(R+2+i)}g(i+1)\end{aligned} \tag{59}$$

The first summation dominates the second in the above equation; the first grows like $\log^2 m$, while the second grows like $m\log m$.

The same pattern emerges at all orders – the dominant contribution can be isolated as :

$$\begin{aligned}\gamma_m^{(i)} &\sim \gamma_1^{(i)} + \sum_{j_1=2}^{m} g(j_1)\sum_{j_2=2}^{j_1} g(j_2)\cdots\sum_{j_i=1}^{j_{i-1}} g(j_i) \\ &\approx \gamma_1^{(i)} + \frac{1}{i!}\left(\sum_{j=2}^{m} g(j)\right)^i\end{aligned} \tag{60}$$

The sum of the dominant contributions remains finite:

$$\gamma_m \sim \exp\left[Q\sum_{j=2}^{m} g(j)\right] \sim \exp\left(-QR\log m\right) \tag{61}$$

and suggests that for large m, γ_m will decay as a power-law with exponent QR.

Motivated by this observation, and recalling that for large m, $A_m \sim 1/m^{R+2}$ (from equation (9)), we suggest the following approximation for B_m, valid for all values of m, not just when m is large:

$$B_m = C\frac{RN_0}{R+2+QR}\prod_{i=1}^{m-1}\frac{i}{R+2+QR+i} \tag{62}$$

where C is a constant that is independent of m. The above expression for B_m is derived by replacing R by $R + QR$ in the denominator of the product that defines A_m (Equation (30)) This is really nothing more than informed guesswork; this is the simplest expression for B_m that recovers a power-law with exponent $R + QR$ for large m and reduces to A_m when $Q = 0$.

In order to determine $F(t)$, the total number of folds at time t, equation (11) has to be solved using the approximate solution (62). First, the a choice has to be made for the constant C – since the equation is an approximation, there is freedom in the choice. One way is to enforce the consistency of equation (56) for $m = 1$:

$$\begin{aligned} \gamma_1^{(1)} = \frac{B_1}{A_1} = C\frac{R+2}{R+2+QR} = 1 + \frac{2}{(R+2)(R+3)} \\ \Longrightarrow C = \left(1 + \frac{2}{(R+2)(R+3)}\right)\left(1 + \frac{QR}{R+2+QR}\right) \end{aligned} \tag{63}$$

As $F(t)$ is directly affected by B_1, it is natural to focus on $m = 1$. Note that for small Q, $C \approx 1 + 2/(R+2)(R+3)$.

Equation (11) can be integrated to give an approximation for $F(t)$:

$$F(t) \approx N_0 + R\left(1 - \frac{QC}{R+2}\right)t \tag{64}$$

Using the identity of Appendix H, the normalized coefficients are given by:

$$p_m = \frac{B_m}{\sum_{m=1}^{\infty} B_m} = \frac{R+1+QR}{R+2+QR}\prod_{i=1}^{m-1}\frac{i}{R+2+QR+i} \tag{65}$$

$$\begin{aligned} F(t) &= N_0 + R\left(1 - \frac{QC}{R+2}\right)t \\ C &= \left(1 + \frac{2}{(R+2)(R+3)}\right)\left(1 + \frac{QR}{R+2+QR}\right) \end{aligned} \tag{66}$$

In the presence of gene deletion, the approximation for $F(t)$ shows linear growth with time at a rate less than R. As expected, a greater rate of gene deletion reduces the growth of $F(t)$. However the approximation predicts that the number of folds will always increase with time, which can be verified by taking the uppermost limit, $Q = 1$. For small Q, the constant C itself can be approximated more simply: $C \approx 1 + 2/(R+2)(R+3)$.

Figure 9 confirms these observations. The approximation for the expected number of folds seems to work quite well and could be useful in trying to infer both R and Q from genomic data. Certainly the impact of gene deletion is easier to identify through $F(t)$ and $G(t)$ than through the shape of the histogram $F(m, t)$.

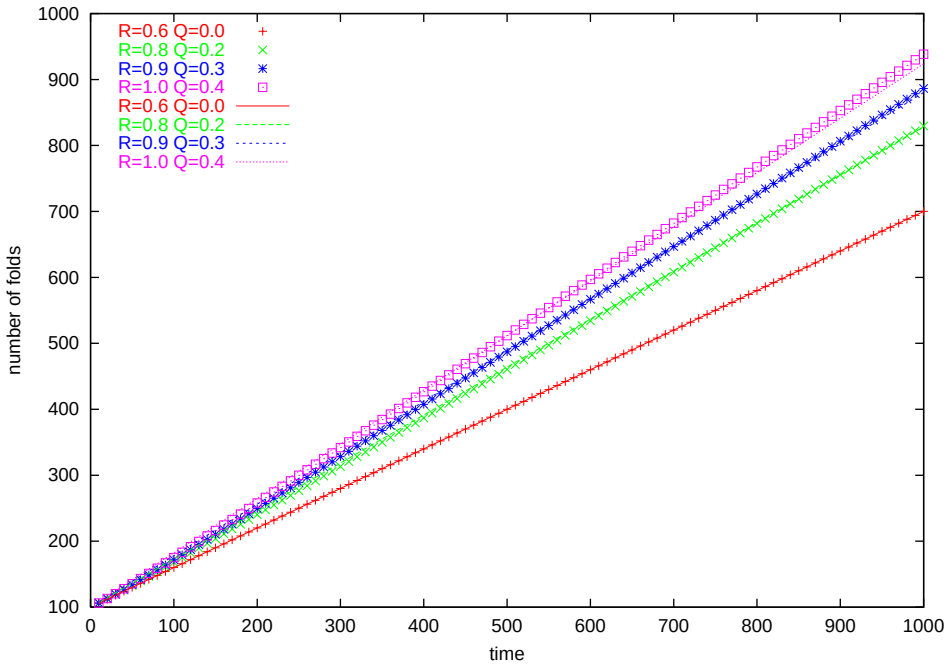

Figure 9: Analytic approximation for the total number of folds compared to numerical results of Figure **??**.