Analytical evolutionary model for protein fold occurrence in genomes, accounting for the effects of gene duplication, deletion, acquisition, and selective pressure

Michael Kamal[1], Nicholas M Luscombe[2], Jiang Qian[3], Mark Gerstein[2,4]

Running title: Analytical model for genome evolution

[1] **Whitehead Institute for Biomedical Research**
Center for Genome Research
Nine Cambridge Center
Cambridge, MA 02142

[2] **Department of Molecular Biophysics and Biochemistry**
and [4] **Department of Computer Science**
Yale University, 266 Whitney Avenue
PO Box 208114
New Haven, CT 06520-8114

[3] **Department of Ophthalmology,**
Johns Hopkins Medical School,
Baltimore, MD 21287

Abstract

Motivation: Global surveys of protein folds in genomes measure the usage of essential molecular parts in different organisms. In a recent survey, we showed that the occurrence of protein folds in 20 completely sequence genomes follow a power-law distribution; i.e. the number of folds (F) with a given genomic occurrence (V) decays as $F(V) = aV^{-b}$, with a few occurring many times and most occurring infrequently. Clearly, such a distribution results from the way in which genomes have evolved into their current states.

Results: Here we develop and discuss a minimal, analytically tractable model to explain these observations. In particular, we demonstrate that (i) stochastic gene duplication and (ii) overall acquisition of new folds are sufficient to accurately replicate the power-law distributions. Furthermore by optimizing the model using genomic data, we gain a quantitative insight into otherwise unattainable data. In particular, as the rate at which genomes acquire new folds is directly related to the power-law exponent -b, we can easily estimate this rate by measuring the gradient of the distribution on a log-log graph.In addition, extensions to the model suggest that gene deletion and selective pressure are important to the fate of individual genes, but do not significantly affect the final power-law distribution. That is, although gene deletion and selective pressure will affect the choice of the most common fold type in an organism, it will not change the overall power-law distribution found across different genomes. Finally, we gain an indication of the initial sizes of genomes, from the starting states of the simulations. We find that the power-law dependence of the fold distribution is independent of the composition of the starting genome.

Availability: Additional data pertaining to this work is found at
http://www.partslist.org/powerlaw

1   Introduction

The power-law behavior is frequently found in many different population distributions. Also referred to as Zipf's law, a well-documented example is the usage of words in text documents (Zipf, 1949). By grouping words that have similar occurrences, it was noted that a small selection such as "the" and "of" are used many times, while most occur infrequently. When the size of each group is plot against its usage, the distribution is described by a power-law function: the number of words (F) with a given occurrence (V) decays with the equation $F = a/V^b$. The distribution is linear when plot on log-log axes, where -b describes the slope. Such distributions are also found for the relative sizes of cities, income levels and the number of papers published by scientists in a field of research (Zipf, 1949).

Significantly, the power-law behavior is also prevalent in many aspects of genomic biology (Luscombe et al., submitted). It is found in the usage of short nucleotide sequences (Mantegna et al., 1994; Konopka & Martindale, 1996; Israeloff et al., 1995; Bonhoeffer et al., 1996; Voss et al., 1996), the populations of gene families (Gerstein, 1997; Huynen & van Nimwegen, 1998), the occurrence of protein superfamilies and folds in genomes (Koonin et al., 2000; Qian et al., 2001) and several biological networks (Jeong et al., 2000; Park et al., 2001; Rzhetsky and Gomez, 2001). The distribution extends even further to the number of distinct protein functions associated with a particular fold, the number of protein-protein interactions that are made by each fold type, and the variations in expression levels between genes present in the yeast genome. These observations have been made in at least 20 prokaryotic and eukaryotic genomes, and so are likely to be universal to most other genomes that are yet to be analyzed. Given the prevalence of this behavior, we suggest that all of these biological distributions

2

arise because of a common mechanism for genomic evolution, primarily by duplicating existing genes to increase the presence of particular types of proteins (Qian et al., 2001).

The current study focuses on the distribution of protein folds in different organisms (Figure 1). Most proteins encoded in a genome have a defined three-dimensional structure that can be classified into distinct protein folds. Although these folds are defined by the topology of the peptide chain, it is possible to determine whether two proteins adopt the same fold by sequence comparison. So even if structures are unavailable for all the genes, we can classify them into equivalent folds by sequence similarity. Using these classifications, one way of representing the contents of a genome is to count the number of times different folds occur and then group together those with similar occurrences (Figure 1). Like word usage, the number of folds (F) with a certain genomic occurrence (V) decays according to the power-law function; we display the distribution for the E. coli genome in Figure 1, and plots for 19 further organisms is available from our supplementary website.

There have been several efforts to understand this nonuniform distribution of protein families. A number of models suggested that the observation of non-uniform population distributions of protein families depends on the "designability" of the protein structure; that is, the relative size of a family depends on the fraction of all sequences that could successfully fold into any particular protein fold (Taverna 2000, Shakhnovich 1998). Others have modelled the occurrence of non-uniform distributions by simulating the evolution of genomes. In the model of Huynen and van Nimwegen (1998), families expand or shrink in size through successive multiplications by a random factor, which represents duplication or deletion events depending on its value. More recently, Yanai et al. (2000) introduced a model in which a genome evolves from a set of precursor genes to a mature size by iterative gene duplications and gradual accumulation of modifications through point mutations. When an individual family member acquires enough random mutations, it breaks away to form a new family.

We recently presented an equally minimal, but more biologically realistic model (Qian et al., 2001). Here genomes evolve through stochastic gene duplications and steady acquisition of new protein folds, either by ab initio creation or horizontal gene transfer (Lawrence & Ochman, 1998; Doolittle, 1999; Ochman et al., 2000; Kidwell, 1993; Doolittle, 1993; de la Cruz & Davies, 2000). Simulations replicated the genomic distributions very accurately, and provided insight into the rate at which different organisms acquired new folds and the origins of a common ancestral genome. Although our work focused on the distribution of protein fold populations, the model applied equally well for other gene classifications such as sequence families, and SCOP superfamilies (Lo Conte, 2000).

The purpose of the current work is two-fold. First, we propose new models based on our previous model by fully incorporating two additional processes in evolution: gene deletion and selective pressure. These major biological processes were beyond the scope of our previous work, and are likely to affect the outcome of our evolutionary model. Second, we provide full analytical and numerical analyses of the original model and its extensions; in doing so we explore the mathematical and biological significance of the model, and explore the relative effects that the different evolutionary processes (gene duplication, acquisition, deletion and selective pressure) have on the final appearance of different genomes. In the previous paper, our results are only based on simulations. In contrast, the analytical approach is also employed in this work.

2 Minimal Model: Gene Duplication and New Fold Acquisition

Suppose that the initial genome consists of N_0 distinct folds at time $t = 0$, i.e. the number of genes equal the number of folds. The growth of the genome in our model occurs by randomly duplicating existing genes, and by incorporating new folds into the genome at a constant rate. Both of these processes are assumed to operate independently and continually over time. We assume that at every instant, all genes are equally likely to be chosen for duplication and that on average, one duplication event happens per unit time. As a result, large folds, i.e. ones that are coded by many genes, are more likely to grow over time than smaller folds. We assume that R new folds of size 1 are always incorporated in to the genome per unit time, i.e. the acquisition of new folds is not stochastic.

Let $F(m,t)$ be the expected number of folds of a given size m at time t. The fold histogram determines both the expected total number of distinct folds $F(t)$ and the expected total number of genes $G(t)$:

$$F(t) = \sum_{m=1}^{\infty} F(m,t)$$
$$G(t) = \sum_{m=1}^{\infty} m F(m,t) \tag{1}$$

Under these growth assumptions, the Markovian dynamics governing $F(m,t)$ are given by:

$$\frac{\partial F(m,t)}{\partial t} = \frac{(m-1)F(m-1,t)}{G(t)} - \frac{mF(m,t)}{G(t)} \quad (m>1)$$
$$\frac{\partial F(1,t)}{\partial t} = R - \frac{F(1,t)}{G(t)} \tag{2}$$

Although duplication occurs at the gene level, it is more convenient mathematically to work directly with the fold histogram $F(m,t)$.

The intuition behind these equations is as follows. If the gene selected for duplication that originally is a member of a fold of size $m-1$, then after duplication that fold will now be a fold of size m, and the population of $F(m-1,t)$ and $F(m,t)$, will decrease and increase, respectively, by one. The probability for this particular gene selection is $\frac{(m-1)F(m-1,t)}{G(t)}$.

These equations ensure the appropriate expected growth rates for the total number of folds. A direct summation of (2) leads to:

$$\frac{\partial F(t)}{\partial t} = \frac{\partial}{\partial t} \sum_{m=1} F(m,t)$$
$$= R \tag{3}$$

and hence: $F(t) = N_0 + Rt$. Similar manipulations show that the expect number of genes also grows as required: $G(t) = N_0 + (R+1)t$. It is important to note that evolution equations enforce the correct overall normalization for the histogram; there is no need to impose normalization conditions separately.

The complete analytical solution for the coupled equations (2) can be found by standard methods. As the calculations are straightforward, but tedious, full details are left to Appendix A.

4

The biological interpretation of the analytical solution is best appreciated by examining two important limiting cases. If there is no acquisition of new genes ($R = 0$), the solution simplifies considerably:

$$F(m,t) = N_0 \phi^{-1} \left(1 - \phi^{-1}\right)^{m-1} \tag{4}$$

where $\phi(t)$ relates the passage of time to the expected number of genes:

$$\phi(t) = \frac{G(t)}{N_0} = 1 + \frac{(R+1)t}{N_0}, \tag{5}$$

Therefore, gene duplication alone leads to an exponential distribution of fold occurrence: $\log F(m,t) = m \log\left(1 - \phi^{-1}\right) + \psi(t)$, with $\psi(t)$ independent of m.

The other revealing limit concerns the behavior for large times ($t \to \infty$) when new genes are acquired at a nonzero rate ($R \neq 0$). The asymptotic limit is given by:

$$F(m,t) \to A_m \phi(t) = A_m \left(1 + \frac{(R+1)t}{N_0}\right) \quad \text{as} \ \ t \to \infty \tag{6}$$

with coefficients A_m that depend only on R and N_0, and not on time:

$$A_m = \frac{RN_0}{R+2} \prod_{i=1}^{m-1} \frac{i}{R+2+i} \tag{7}$$

Consequently, the probability distribution of fold sizes, i.e. the normalized histogram, is determined by solely by the A_m:

$$p(m,t) = \frac{F(m,t)}{\sum_i F(i,t)} \to \frac{A_m}{\sum_i A_i} = \frac{R+1}{R+2} \prod_{i=1}^{m-1} \frac{i}{R+2+i} \tag{8}$$

and furthermore this asymptotic probability distribution depends only on R – the dependence on initial cluster size N_0 is removed by the normalization.

An examination of the the leading large m behavior of A_m reveals that

$$\log A_m \sim -(R+2) \log m \tag{9}$$

Therefore, for large m, the terminal probability distribution (8) resembles a power-law with exponent $R + 2$. For small m, the coefficients decrease less rapidly with m and do not resemble power-law dependence. This observation is relevant for estimating R from empirical data or even numerical results.

It is also worth pointing that a power-law distribution that decays too slowly will lead to an infinite expected number of genes. A power-law distribution will that holds asymptotically for large m: $N(m) \sim 1/m^\alpha$ has to be described by an exponent $\alpha > 2$ for the sum $G(t) = \sum_m^\infty m N(m)$ to converge. The asymptotic limit of the exact solution, a power-law with exponent $R + 2$, satisfies this condition.

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, the analytic solution confirms what would be expected intuitively: the histogram behavior is dominated by duplication events involving the initial N_0 genes. To characterize the "crossover" behavior of the solution from the exponential to approximate power-law regime we have calculated the similarity of the exact probability distribution at different times to both the best fitting exponential distribution and to the limiting asymptotic distribution (8). The difference between any two probability distributions is measured by the sum of squared differences (the standard L^2 metric).

5

We have characterized the crossover time T_c for a range of values for both R and N_0 and find that the crossover time displays two distinct regimes. Within each regime it is approximately inversely proportional to R and directly proportional N_0: $T_c \sim N_0/R$, with a different proportionality constant for each regime. Details of this analysis can be found in Appendix B. The numerical results indicate that crossover occurs roughly when the number of new fold introductions: RT_c, becomes comparable to the initial genome size N_0, as might be expected intuitively.

$$\boxed{\text{THE PLOT}}$$

Figure 1: Fold distribution for E. coli

$$\boxed{\text{THE PLOT}}$$

Figure 2: Three Models: (a) Minimal, (b) Gene deletion and (c) Selective pressure

So far, we have assumed that the starting genome contains just one copy of each fold. In fact, it is reasonable to expect the initial genome to have several copies of particular fold types (for example those involved in protein synthesis) when the evolutionary process described by the model was initiated. By definition, genomes in our model have a comparatively small starting state, and so the difference between the most and least common folds would be minimal, i.e. some occurring three or four times at most. However, it is nonetheless of interest to investigate the effect that the appearance of the initial genome would have on the final distribution.

The solution we have derived for a particular initial genomic configuration – N_0 distinct folds consisting of one gene – can be extended to describe the evolution of an arbitrary initial fold distribution $N_{init}(m)$ that is made up of N_0 genes: $\sum_m m N_{init}(m) = N_0$. The solution is similar to the special initial condition of N_0 distinct folds and is presented in detail in Appendix C.

One important conclusion may be drawn from the generalized model: all initial distributions ultimately lead to the the same limiting distribution determined by the A_m. Just as before, the dependence on the initial fold distribution $N_{init}(m)$ decays with time, leading to the same asymptotic distribution as was found for an initial distribution of N_0 folds of size 1 in (9), reflecting the dominance of fold introduction over gene duplication for large times. Of course, the details of how and when the crossover happens will depend on the particular form of $N_{init}(m)$.

3 Extended Model: Including the Effects of Random Gene Deletion

Gene deletion is a major factor in evolution and is discussed in detail by Qian et al. (2001). In this section we incorporate an additional parameter, Q, that represents gene deletion.

The most natural extension of (2) that accounts for allows random gene deletion at rate Q would be the following:

$$\frac{\partial F(m,t)}{\partial t} = \frac{(m-1)F(m-1,t)}{G(t)} - \frac{mF(m,t)}{G(t)} + Q\frac{(m+1)F(m+1,t)}{G(t)} - Q\frac{mF(m,t)}{G(t)} \quad (m > 1)$$

$$\frac{\partial F(1,t)}{\partial t} = R - (1+Q)\frac{F(1,t)}{G(t)} + Q\frac{F(2,t)}{G(t)}$$

$$(10)$$

The terms proportional to Q encode the dynamics for gene deletion, which are very similar to gene duplication: on average, Q deletions occur for every duplication event and the gene to be deleted is chosen randomly from all the genes in the genome. In this way the population of a given bin m can either decrease due to gene deletion if the gene to be deleted is from bin m itself, or it can increase as a result of a deletion in the neighboring bin $m + 1$.

In this extended model, gene growth occurs at the uniform rate one would expect: $G(t) = N_0 + (1 + R - Q)t$. In contrast, the behavior of the expected number of folds is more complicated:

$$\frac{\partial F(t)}{\partial t} = \frac{\partial}{\partial t} \sum_{m=1}^{\infty} F(m,t)$$

$$= R - Q\frac{F(1,t)}{G(t)}$$

$$(11)$$

7

Folds of size 1 that are deleted disappear from the genome so $F(t)$ depends explicitly on the population of $F(1,t)$; unlike the $Q = 0$ case, the dynamics of $F(t)$ can not be determined without knowing the full solution to (10).

The extended is much more complicated mathematically, primarily because the difference equations are now second order. In the duplication/acquisition model, the behavior of larger folds depends only on the behavior of smaller folds, so the full solution can be constructed inductively starting from the solution for $m = 1$. With gene deletion operating as well, the dynamics of different fold sizes are coupled together. In many respects, these dynamics are like those describing diffusion phenomena; when $Q = 0$ the genome exhibits growth due to drift, or directed movement alone, while nonzero Q introduces diffusive, or non-directional movement as well.

3.1 Analytic Results

We were able to derive a full analytical solution only in the absence of any new fold introduction: $R = 0$. In this case, only stochastic gene deletion and duplication operate. We will restrict our discussion to when gene duplication occurs at a higher rate than gene deletion, which requires $0 < Q < 1$, so the genome will still grow in size, at least in terms of number of genes: $G(t) = N_0 + (1 - Q)t$. Note that since $R = 0$, equation (11) shows that the number of folds will actually decrease with time. Losing folds while gaining genes is possible if the larger folds make up for the loss of genes from smaller folds.

An analytic solution exists for an initial distribution of N_0 different folds of size 1 and is worked out in detail in Appendix D. Once again, the distribution is exponential. Figure 3 shows histograms $F(m,t)$ corresponding to three values of Q and a fixed time.

Remarkably, the normalized distribution of fold size (the probability distribution) is independent of the gene deletion rate Q:

$$
\begin{aligned}
p(m,t) &= \frac{F(m,t)}{\sum_{i=1}^{\infty} F(i,t)} \\
&= \frac{N_0}{N_0 + t} \left[\frac{t}{N_0 + t} \right]^{m-1}
\end{aligned}
\tag{12}
$$

Hence gene deletion does not affect the shape of the distribution at all when $R = 0$, only the overall normalization is changed. This can also be seen directly from Figure 3.

Although an exact analytical solution does not seem possible for arbitrary R and Q, it is nonetheless possible to derive analytic expressions for the higher moments of the fold distribution. Appendix E discusses how this is done and particular, includes an expression for the second moment that will prove useful when fitting the model to genomic data.

3.2 Numerical Results for Nonzero R and $Q < 1$

Numerical solution of (10) reveals for that large times, the normalized histograms of fold size approach a time-invariant limit that depends solely on R and Q. Figure 4(a) shows the probability distributions for a fixed rate of new fold acquisition, $R = 1.0$, and increasing rates of gene deletion: $Q = 0, 0.2, 0.4, 0.6$. The power-law character of the distributions is retained even for large values of Q. Quite reasonably, higher rates of gene deletion encourages the dominance of smaller folds, leading to a more rapid decline of $p(m)$ with fold size m. Common folds require repeated gene duplication and an avoidance of gene deletion events to proliferate. As the probability of avoidance is
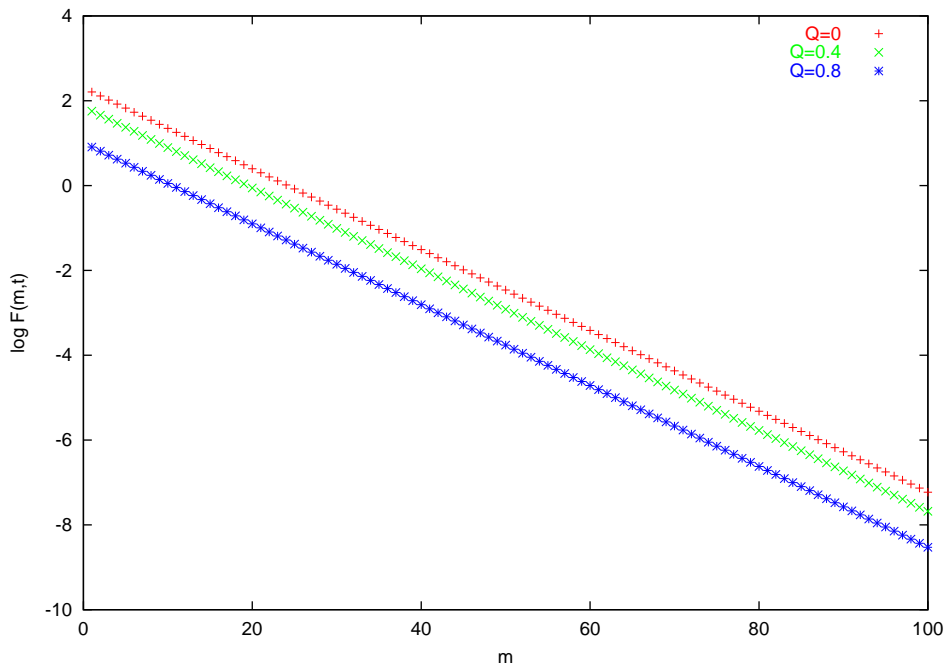
Figure 3: Fold histogram $F(m, t)$ for $N_0 = 100$ and $t = 1000$ plotted for $Q = 0, 0.4, 0.8$ and $R = 0$. Note that $logF(m, t)$ is plotted against m to highlight the exponential dependence.

proportional to $1 - Q$, the probability of multiple avoidance is suppressed as a power of $1 - Q$.

Figure 5(a) explores the effect of deletion when the overall gene growth rate is kept constant: $1 + R - Q = 1.6$. In this way, we can contrast the effects of deletion and fold acquisition in a controlled manner. Note that a commensurate increase in R does not overcome an increase in Q, as large folds are suppressed more than small folds. This means that the exponent that best describes the power-law decay is not merely a function of of $R - Q$.

On the other hand, the effect of gene deletion is not dramatic; not only is similarity to a power-law retained the actual change in exponent is not large. Even for fold of large size, there isn't much difference between the curves even for a fairly large gene deletion rate. In practice, this makes it difficult to estimate Q statistically from the shape of fold histograms derived empirically from genomic data. While the effective gene introduction rate: $1 + R - Q$, should be easy to deduce from the data, an identification of Q itself from the rate of decay would require reliable occurrence data for very large folds.

When there is no gene deletion, the expected number of folds increases linearly with time at rate R. Equation (41) suggests gene deletion will lead to a less simple time dependence for $F(t)$. Perhaps surprisingly, $F(t)$ remains, to a good approximation, linear in time, with a slope that is no longer R, as can be seen in the numerical results of Figure 4(b). Here $F(t)$ is plotted for fixed $R = 1.0$ and different values of the gene deletion rate: $Q = 0.0, 0.2, 0.4, 0.6$. In fact, the slope in each of these cases is less than R and decreases with Q, which is consistent with the analytic solution for $F(t)$ when $R = 0$, derived derived in Appendix D (see equation (41).

If again we choose parameters that fix the growth rate for the expected number of genes $(1 + R - Q)$, a commensurate increase in both R and Q leads to a greater

9
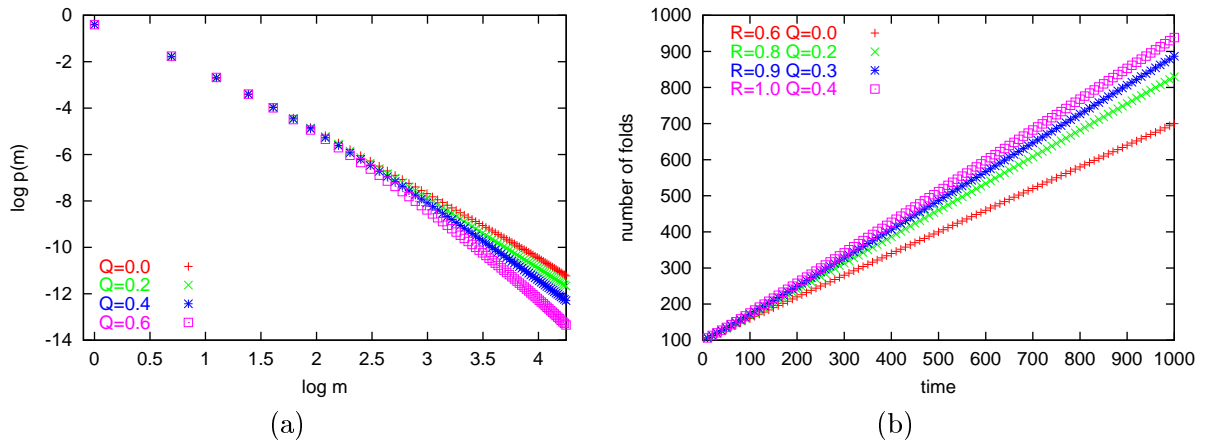
(a)   (b)

Figure 4: (a) Normalized large-time limiting fold distribution and (b) the total number of folds as a function of time when $R = 1.0$ and $Q = 0, 0.2, 0.4, 0.6$. Logarithm of $p(m)$ is plotted against logarithm of m

increase in the expected number of folds, as can be seen in Figure 5(b). This is entirely reasonable: in our model, the new folds that are continually acquired at rate R are all distinct, so a genome with large R and Q will end up with many small folds, each coded by only a few genes. In contrast, a genome with small R and Q will lead to fewer but larger folds.

3.3   Analytic Approximation Based on Perturbation Theory

The numerical results show that gene deletion, even for fairly large values of Q does not dramatically change the growth pattern of the genome, certainly qualitatively and to some extent, even quantitatively. Moreover, the analytic results from section 3.1 showed that gene deletion is remarkably benign: in the absence of new gene acquisition, but with gene duplication operating, gene deletion does not change the probability distribution of fold occurrences, but does change expected total number of folds in the genome.

This suggests that it should be possible to capture the effects of gene deletion perturbatively by constructing an approximation around the $Q = 0$, $R > 0$ solution as an expansion in powers of Q. The perturbation expansion has to be handled carefully since a naive expansion, one that considers contributions only up to some finite power of Q, will not converge for all fold sizes m. The failure of conventional perturbation theory is explored in Appendix F.

To go beyond naive perturbation theory, we have adopted the following approach: (1) the dominant contribution at every order (or power) of Q is identified, (2) the dominant contribution is approximated, and (3) the resulting new infinite series in Q is summed exactly to arrive at an approximate solution that remains finite for all Q and m. The details are presented in Appendix F. Although not rigorous, this type of rescue or augmentation of perturbation theory is practiced routinely and often quite successfully on a variety physical models, such as models of phase transitions from statistical physics (Itzykson, 1989).

This approach leads to the following approximation for the limiting fold distribution:

$$p_m = \frac{R + 1 + QR}{R + 2 + QR} \prod_{i=1}^{m-1} \frac{i}{R + 2 + QR + i} \qquad (13)$$
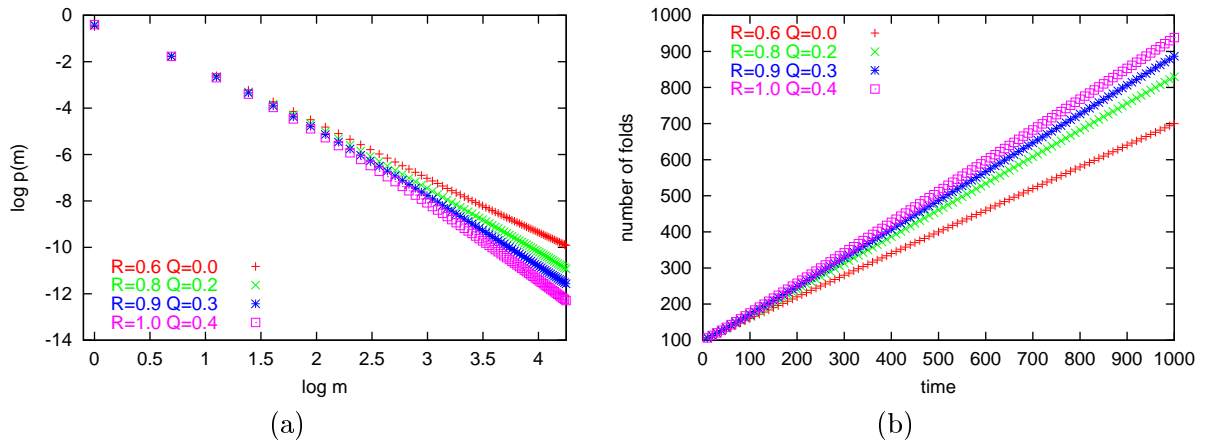
10

(a)   (b)

Figure 5: (a) Normalized large-time fold distribution and (b) the total number of folds as a function of time for fixed overall gene growth: $1 + R - Q = 1.6$ and $Q = 0, 0.2, 0.3, 0.4$. Logarithm of $p(m)$ is plotted against logarithm of m

Note that the approximation includes as a special case the exact distribution derived previously for $Q = 0$ (8). In fact, the approximate distribution for Q nonzero is obtained from the exact solution for $Q = 0$ by the substituting $R \rightarrow R + QR$. This correspondence also makes it clear that for large m, the $Q \neq 0$ probability distribution will resemble a power law with exponent $R + QR + 2$, just as $Q = 0$ distribution approached a power-law with exponent $R + 2$.

The true test of the effectiveness of the approximation rests with a comparison to the numerical results, which is done in Figure 6. There seems to be good qualitative agreement, and fairly good quantitative agreement as well, even for $Q = 0.4$. As expected from the nature of the approximation, there is better agreement for large m in all cases. An approximation for the expected number of folds $F(t)$ within the same framework is given in Appendix F.

4   The Effects of Selection Pressure

Selective pressure plays an important role in the evolution. It is well known that different genes have different duplication rate due to the selective pressure (Patthy, 1999). So far we have assumed that when genes are duplicated, or deleted, the target gene is chosen with equal probability from all the genes in the genome. A more realistic model would of course allow for favoritism in the selection process: presumably, genes that are useful or necessary are less likely to be deleted and perhaps more likely to be duplicated than genes that are less important. Note, however, that our model is not a differential survival model.

We explore the effects of selection pressure by extending the minimal model to allow for different duplication rates among genes. Suppose now that genes are not only identified with particular folds but also by their duplication types. For simplicity, assume that there are only two 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.
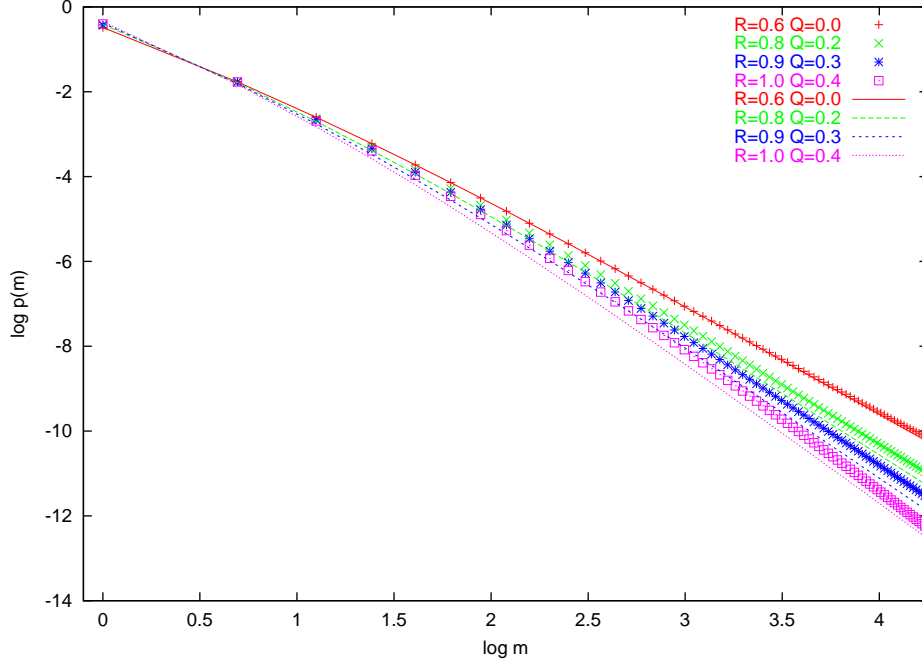
11

Figure 6: Analytic approximation, shown using solid lines, for normalized terminal fold distribution, compared to the numerical results of Figure 4.

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 $G_\gamma(t)$:

$$G_\gamma(t) = G_A(t) + \gamma G_B(t) \tag{14}$$

The evolution equations that extend (2) are:

$$
\begin{aligned}
\frac{\partial F_A(m,t)}{\partial t} &= \frac{(m-1)F_A(m-1,t)}{G_\gamma(t)} - \frac{mF_A(m,t)}{G_\gamma(t)} \quad (m>1) \\
\frac{\partial F_A(1,t)}{\partial t} &= R_A - \frac{F_A(1,t)}{G_\gamma(t)} \\
\frac{\partial F_B(m,t)}{\partial t} &= \gamma\frac{(m-1)F_B(m-1,t)}{G_\gamma(t)} - \gamma\frac{mF_B(m,t)}{G_\gamma(t)} \quad (m>1) \\
\frac{\partial F_B(1,t)}{\partial t} &= R_B - \gamma\frac{F_B(1,t)}{G_\gamma(t)}
\end{aligned}
\tag{15}
$$

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.

The equations for the total number of genes of both types follow from the full dynamics (68) and are given in Appendix (G). These confirm that the overall duplication rate is still one gene per unit time.

Once again, analytical solutions are possible for the two special parameter values addressed previously: (1) when there is no introduction of new folds, so $R_A = R_B = 0$;

12

and (2) the limiting distribution when $t \to \infty$. When there is no introduction of new folds, a simple extension of the methodology 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. The full histogram is consequently a sum of exponential distributions:

$$
\begin{aligned}
p(m, t) &= \frac{F_A(m, t) + F_B(m, t)}{\sum_i F_A(i, t) + F_B(i, t)} \\
&= \frac{N_0^A}{N_0^A + N_0^B} e^{-u} \left[1 - e^{-u}\right]^{m-1} + \frac{N_0^B}{N_0^A + N_0^B} e^{-\gamma u} \left[1 - e^{-\gamma u}\right]^{m-1}
\end{aligned} \tag{16}
$$

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 a rescaled time variable related to $G_\gamma(t)$; the exact form of the dependence appears in Appendix G but is unimportant for the present discussion.

Of greater interest is the other special case: the ultimate evolutionary fate of the genome. The analytic behavior for large times is much easier to derive than an exact solution itself. 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 (see Appendix G for details).

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 (8), 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 normalized probability distribution corresponding to this limit is given by:

$$
\begin{aligned}
p(m, t) &= \frac{C_m^A + C_m^B}{\sum_i C_i^A + C_i^B} \\
&= \frac{C_\gamma}{C_\gamma + 1} \frac{R_A}{R_A + R_B} \prod_{i=1}^{m-1} \frac{i}{C_\gamma + i + 1} + \frac{C_\gamma}{C_\gamma + \gamma} \frac{R_B}{R_A + R_B} \prod_{i=1}^{m-1} \frac{i\gamma}{C_\gamma + \gamma(i + 1)}
\end{aligned} \tag{17}
$$

The important conclusion to be drawn from (17) is that powerlaw-like distributions describe the ultimate fate of the genome even when there are different rates of gene duplication. The probability distribution is the sum of two powerlaw-like distributions, each similar to the powerlaw-like distributions of the minimal model, but characterized by its own effective exponent. Figure 7 shows a comparison of the predicted distribution and numerical results when $R_A = R_B = 0.5$ for $\gamma = 1$, which is corresponds to the minimal model, and $\gamma = 10$, so type "B" genes are ten times more likely to be selected for duplication. The two distributions are remarkably close to each other, even when there is an order of magnitude difference between the relative duplication rates of type "B" genes. We have found that the parameter γ has much less of an effect than differences between the gene introduction rates R_A and R_B.

We have also briefly considered the case of more than two duplication types. When there is no introduction of new folds into the genome equation (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 four duplication types appears in Appendix G.

5   Fitting the Models to Genomic Data

Clearly, of greatest interest is to observe how our model compares with the genomic distributions. We start with the minimal model, for which we require estimates for the parameters t, N_0 and R for each organism.
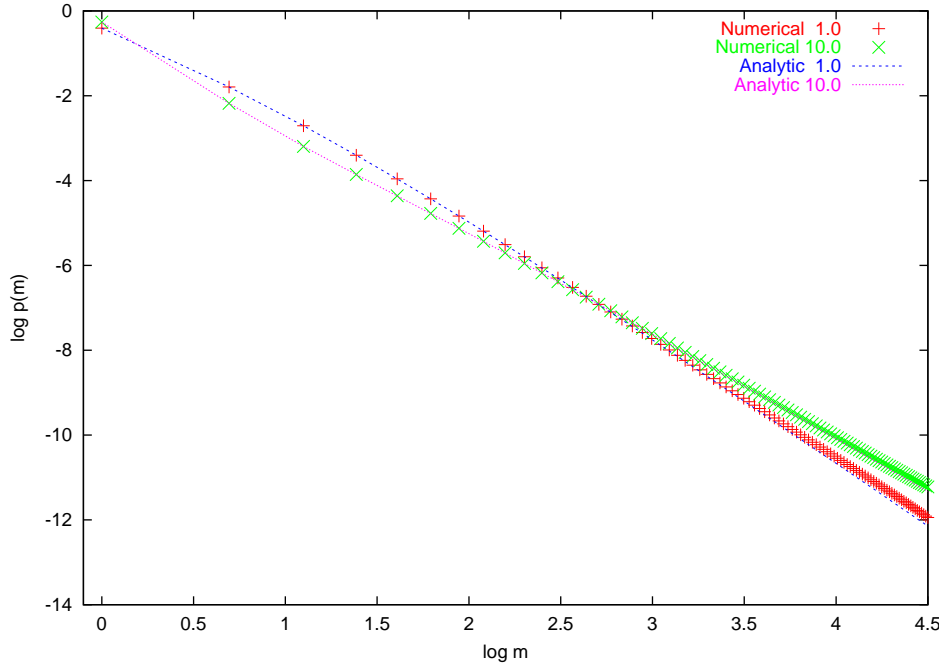
Figure 7: Large time limit for the fold probability distribution for $\gamma = 1$ and $\gamma = 10$. Numerical results are plotted as symbols; analytic results from Equation (17) as lines

Fitting the minimal model requires estimating three parameters: t, N_0 and R. We have determined these parameters separately for each organism by insisting that the minimal model match the number of folds: F, the number of genes: G, and the second moment of the actual fold histogram: H_2, to those predicted by the minimal model. The fitting procedure is greatly simplified by the linear relation that exists between the variables (N_0, t) and (F, G):

$$N_0 = (R + 1)F - RG$$
$$t = G - F \tag{18}$$

The estimation of R is aided by recasting the expression for H_2 (Eq. 47 in Appendix E) so that t no longer appears explicitly. Instead, the second moment can be expressed so that it depends directly on F, G and the unknown R:

$$\frac{H_2}{G} = \frac{R + 1}{R - 1} - \frac{2}{R - 1} \left[\frac{\frac{G}{F}}{R + 1 - R\frac{G}{F}} \right]^{\frac{1 - R}{1 + R}} \tag{19}$$

This equation is well behaved and can be easily solved numerically. What threatened to be a coupled, nonlinear three dimensional estimation problem is actually nothing more than a single nonlinear equation and two linear equations. We have verified that this fitting procedure accurately recovers parameters values from distributions generated both numerically and from the exact solution.

The results appear in Table 1. As a measure of the quality of the fit, we also report the mismatch of between the third moment predicted by the minimal model and observed in the data, as a percentage of the observed value; a positive value indicates that the model moment is larger. Plots of the actual fits appear in Figure 8.

14

Genome	Genes	Folds	G/F	t	N_0	R	Mismatch of third moment (%)
M. genitalium	481	200	2.40	281	7	0.69	9.9
M. pneumonia	688	277	2.49	411	15	0.637	3.6
R. prowazeki	834	322	2.59	512	26	0.576	-13.1
C. trachomatis	894	344	2.60	550	29	0.574	-8.0
T. pallidum	1031	367	2.81	664	31	0.505	-14.1
C. pnemoniae	1052	390	2.70	662	34	0.538	-10.2
A. aeolicus	1522	357	4.26	1165	68	0.249	-3.0
H. pylori	1553	477	3.26	1076	52	0.395	0.6
B. burgdorferi	1638	559	2.93	1079	13	0.506	4.5
H. influenzae	1709	457	3.74	1252	70	0.31	0.9
M. jannaschii	1715	358	4.79	1357	34	0.239	3.8
M. thermoautotrophicum	1869	374	5.00	1495	35	0.227	-10.5
P. horikoshii	2064	450	4.59	1614	91	0.223	-5.7
A. fulgidus	2420	419	5.78	2001	72	0.173	-3.2
Synechocystis sp.	3169	558	5.68	2611	108	0.172	0.3
M. tuberculosis	3918	491	7.98	3427	118	0.109	-2.2
B. subtilis	4100	584	7.02	3516	153	0.123	-12.4
E. coli	4289	610	7.04	3679	141	0.127	-5.2
S. cerevisiae	6269	575	10.9	5694	128	0.078	2.8
C. elegans	19099	605	31.6	18494	120	0.026	-18.4

Table 1: Fit of the minimal model using genomic data from 20 organisms

The parameter values are in fact very similar to those obtained in our previous work. The mismatch values range -13.1% to 9.9% and indicate that the distributions resulting from our model closely resembles the genomic distribution.

Our attempts to fit the models that included gene deletion were not that informative. This is partly because, as we have seen already, the gene deletion parameter Q does not have a dramatic effect on the shape of the distribution. We had difficulties even trying to fit distributions generated numerically from the extended model. Unlike the equations describing the minimal model, these coupled equations are also nonlinear. Furthermore, since there is no exact analytic expression for $F(t)$, one of the variables itself has to be calculated numerically (We found that our analytic approximation for the number of folds given in Appendix F was not accurate enough to carry out the root-finding). We have found that naive multidimensional root-finding algorithm are either unable to distinguish between many approximate solutions, or find no solution at all at with increased sensitivity. The same difficulties were encountered in trying to discern evidence for selection pressure – there was too little dependence on the selection parameter γ to allow reliable estimation.

6 Conclusions

Here we propose two new models based on our previous model by fully incorporating two major processes in evolution: gene deletion and selective pressure. Both mathematically and biologically, including these effects are not slight. Mathematically, the derivations clearly show they are not trivial. Biologically, these effects provide a much
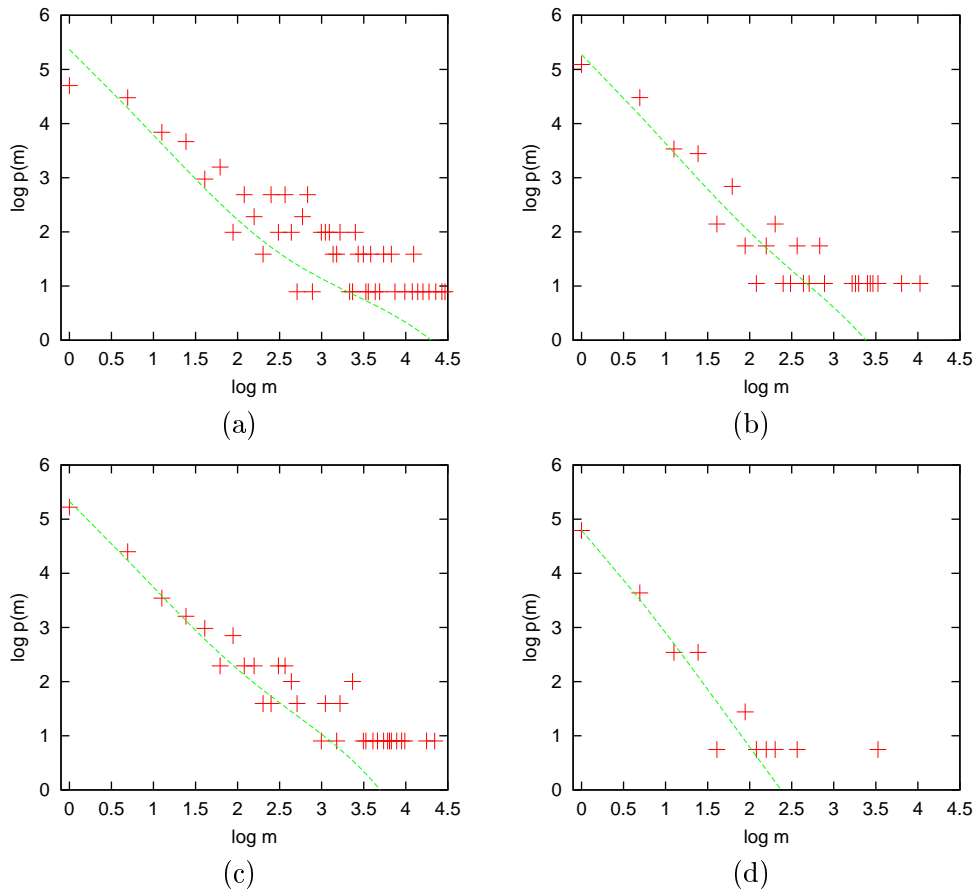
Figure 8: Minimal model fits for (a) C. elegans, (b) A. fulgidus, (c) M. tuberculosis and (d) M. genitalium using parameters from Table 1

more realistic model for genomic evolution than has been presented in any previous publications (Huynen and van Nimwegen, Yanai et al. Qian et al). Furthermore, we provide analytical and numerical analyses of the original model and its extensions to explore the mathematical and biological significance of the models and to demonstrate the effects that the different evolutionary processes (gene duplication, acquisition, deletion and selective pressure) have on the final appearance of different genomes.

The field of the power-law distributions is controversial (Mantegna et al. 1994; Gerstein 1997; Huynen and van Nimwegen 1998; Jeong et al. 2000; Koonin et al. 2000; Park et al. 2001; Konopka and Martindale 1995; Bonhoeffer et al. 1996b; Bonhoeffer et al. 1996a; Israeloff et al. 1996; Martindale and Konopka 1996; Voss 1996). A number of fitting functions other than the power law were proposed to explain the observation (Govindarajan 1999, Martindale 1996). Our argument is that the question of which fitting function is the best should not be the central problem, because one can always find a function with more parameters fits the observation better than others (Luscombe et al., submitted). Instead, we think biologically meaningful models are more helpful for understanding the origin of distribution and the analytical and numerical solutions shown in this work are vital for explaining the observation and further predicting the behaviour of the system.

The full analytical solution to this basic model revealed new facts that were unattain-

able from simulations only. As observed previously, gene duplication alone gives rise to an exponential distribution. However, the combined effect of duplication and acquisition changes the nature of genomic growth dramatically; beyond a sufficient length of evolutionary time, the fold distribution undergoes a transition from the exponential form, to a time-invariant limiting distribution that resembles a power law. The rate of fold acquisition (R) and the size of the initial genome (N0) have distinct effects. Firstly, the cross-over time from the exponential to power-law phases is proportional to N0 and approximately inversely proportional to R. This implies that the transition occurs when the number of new fold acquired becomes comparable to the initial size of the genome. Secondly, the decay rate of the power-law distribution i.e. the slope on a log-log plot is equal to R + 2 for large fold sizes. In fact, the final appearance of the distribution is independent of N0, and is unaffected by the nature of the fold distribution in the starting genome. We find that the decay rate of the power-law distribution i.e. the slope on a log-log plot is equal to R + 2 for large fold sizes.

Note that we take R as a constant, and we regard this as the average rate of fold acquisition throughout the entire course of evolution. In reality, the value of R is likely to vary with time owing to a number of factors such as the decrease of available new protein folds. Further effects might be the increasing difficulty in horizontally transferring genes as the organism becomes more complex. These effects would generally lead to a decrease in rate of fold acquisition with time and this is perhaps reflected in the lower values of R for larger genomes.

We also studied extended models that fully incorporate the effects of random gene deletion and selective pressure. Gene deletion, represented by the parameter Q, does not significantly alter the qualitative behaviour found in the minimal model. The analytic solution showed that when there is no fold acquisition (R = 0), the distribution is again exponential and surprisingly, completely independent of Q. For cases where there is fold acquisition (R ¿ 0), gene deletion had two main effects: firstly the final genome contained fewer fold types, and secondly all fold groups had smaller occurrences. Unsurprisingly, the extent of these effects was dependent on the size of Q. The final distribution nonetheless remains close to a power law, with a decay rate of R + 2 + QR.

The effects of selective pressure were incorporated into the minimal model by introducing favouritism into the gene selection process. This was done by having two groups of genes, one with a higher probability of selection than the other. In this case, the two sets of genes effectively evolve with two distributions, each undergoing a transition from the exponential to power-law phases. Therefore, the final fold distribution is the sum of two power-law distributions, which in fact still closely resembles the distribution when no selective pressure is present. This is true even for large differences in duplication probabilities between the two sets of genes. More generally, we could imagine an array of finer differences in duplication probabilities representing the full range of selection pressures for genes of distinct biological functions. For this, we conjecture that selective pressure, at least when modelled as a duplication bias, will lead to folds that co-exist and compete for prominence in the genome, each undergoing separate, but linked distributional transformation.

We compared our minimal model compares with the genomic data by fitting parameter values. Figure 13 and the mismatch values in Table 1 show, the fits between the model and genomic data are good. As discussed in our earlier work, the parameters can be interpreted in a biologically meaningful way (Qian et al., 2001). We did not use the new models for simulating biological data for two reasons: (1) they do not greatly affect the final appearance of the distribution; (2) if we would be trying to fit a model with three additional free parameters, this would detract from the main results of the paper.

In conclusion, although our model considers a few of the many important processes underlying evolution, it is significant that a simplistic model based on gene duplication and fold acquisition leads to distributions close to those observed in genomic data. The current genomes provide only a snapshot in evolutionary time, but through our model, we gain a glimpse into the biological processes that are most important. Furthermore, by estimating parameter values, we obtain quantitative estimates such as the rate of gene acquisition, which would be otherwise unattainable. Interesting expansions to our model in future may include allowing parameter values to vary during the course of evolution, and modelling the evolution of different genomes simultaneously and simulating their divergence into different organisms.

References

Bonhoeffer, S., Herz, A., Boerlijst, M., Nee, S., Nowak, M. & May, R. (1996) No signs of hidden language in noncoding DNA. *Phys Rev Lett,* **76**, 1977.

de la Cruz, I. & Davies, I. (2000) Horizontal gene transfer and the origin of species: lessons from bacteria. *Trends Microbiol,* **8** (3), 128–133.

Doolittle, W. (1998) You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends Genet,* **14** (8), 307–11.

Doolittle, W. (1999) Phylogenetic classification and the universal tree. *Science,* **284** (5423), 2124–9.

E.I.Shakhnovich (1998) Protein design: a perspective from simple tractable models. *Fold Des.,* **3** (3).

Gerstein, M. (1997) A structural census of genomes: comparing bacterial, eukaryotic and archaeal genomes in terms of protein structure. *J. Mol. Biol,* **274** (4).

Huynen, M. & van Nimwegen, E. (1998) The frequency distribution of gene family sizes in complete genomes. *Mol. Biol. Evol.,* **15** (5), 583–9.

Israeloff, N. E., Kagalenko, M. & Chan, K. (1996) Can Zipf distinguish language from noise in noncoding DNA? *Phys Rev Lett,* **76**, 1976.

Itzykson, C. & Drouffe, J. (1989) *Statistical Field Theory, Vols. 1 and 2.* Cambridge University Press, Cambridge.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. & Barabasi, A. L. (2000) The large-scale organization of metabolic networks. *Nature,* **407**, 651–4.

Kidwell, M. (1993) Lateral transfer in natural populations of eukaryotes. *Annu Rev Genet,* **27**, 235–56.

Konopka, A. K. & Martindale, C. (1995) Noncoding DNA, Zipf's law, and language. *Science,* **268**, 789.

Koonin, E., Wolf, Y. & Aravind, L. (2000) Protein fold recognition using sequence profiles and its application in structural genomics. *Adv Protein Chem,* **54**, 245–75.

Lawrence, J. & Ochman, H. (1998) Molecular archaeology of the Escherichia coli genome. *Proc Natl Acad Sci USA,* **95** (16), 9413–7.

LoConte, L., Ailey, B., Hubbard, T., Brenner, S., Murzin, A. & Chothia, C. (2000) SCOP: a structural classification of proteins database. *Nucleic Acids Res,* **28** (1), 257–9.

Luscombe, N., Qian, J., Johnson, T. & Gerstein, M. (Submitted) Power-law behaviour applies to a wide variety of genomic properties. *Trends Genet.,* .

Mantegna, R., Buldyrev, S., Goldberger, A., Havlin, S., Peng, C., Simons, M. & Stanley, H. (1994) Linguistic features of noncoding DNA sequences. *Phys Rev Lett,* **73** (23), 3169–72.

Ochman, H., Lawrence, J. & Groisman, E. (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature,* **405** (6784).

Park, J., Lappe, M. & Teichmann, S. (2001) Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. *J Mol Biol,* **307** (3), 929–38.

Patthy, L. (1999) *Protein Evolution.* Blackwell Science, London.

Qian, J., Luscombe, N. & Gerstein, M. (2001) Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. *J Mol. Biol.,* **313**, 673–681.

Rzhetsky, A. & Gomez, S. (2001) Birth of scale-free molecular networks and the number of distinct dna and protein domains per genome. *Bioinformatics,* **17** (10), 988–96.

Simmons, G. (1994) *Differential Equations with Applications and Historical Notes.* Prentice-Hall, Englewood Cliffs, N.J.

Taverna, D. & Goldstein, R. (2000) The distribution of structures in evolving protein populations. *Biopolymers,* **53** (1), 1–8.

Voss, R. F. (1996) Comment on "Linguistic features of noncoding DNA sequences". *Phys Rev Lett,* **76**, 1978.

Yanai, I., Camacho, C., & DeLisi, C. (2000) Predictions of gene family distributions in microbial genomes: evolution by gene duplication and modification. *Phys Rev Lett,* **85** (12), 2641–4.

Zipf, G. (1949) *Human Behaviour and the Principle of Least Effort.* Addison-Wesley, Cambridge, MA.

A   Appendix: Analytic Solution of the Minimal Model

It helps to introduce a new parameterization of time:

$$u = \log \phi(t) = \log \left(1 + \frac{(R+1)t}{N_0} \right) \tag{20}$$

with associated derivative:

$$\frac{\partial}{\partial t} = \frac{R+1}{N_0 e^u} \frac{\partial}{\partial u} \tag{21}$$

With this definition, $u = 0$ corresponds to $t = 0$.

This new variable helps rid the differential equations (2) of explicit time dependence:

$$\frac{\partial F(m,u)}{\partial u} + \frac{mF(m,u)}{R+1} = \frac{(m-1)F(m-1,u)}{R+1} \quad (m > 1)$$

$$\frac{\partial F(1,u)}{\partial u} + \frac{F(1,u)}{R+1} = \frac{N_0 R}{R+1} e^u \tag{22}$$

19

Note that the equation the special bin $F(1, u)$ does not depend on any other $F(m, u)$, so it can be solved separately. Once it is known, the solution for any other m can be found by successive integration:

$$F(m+1, u) = \exp\left(-\frac{m+1}{R+1}u\right) \int_0^u \frac{dv}{R+1} \, mF(m, v) \exp\left(\frac{m+1}{R+1}v\right) \quad \text{for } m+1 = 2, 3, \cdots$$

(23)

The solution for m serves as a "source" for $m+1$. The relation (23) follows by multiplying both sides of (2) by $\exp\left(\frac{m+1}{R+1}u\right)$ and integrating. Note that this solution ensures that $F(m > 1, t = 0) = 0$, so the initial conditions are automatically satisfied. Our method of solving the differential equation is elementary and standard, see (Simmons, 1994) for more details.

The solution for $m = 1$ can be found in the same way:

$$\frac{\partial}{\partial u}\left[\exp\left(\frac{u}{R+1}\right)F(1, u)\right] = \exp\left(\frac{u}{R+1} + u\right)\frac{N_0 R}{R+1}$$

(24)

$$F(1, u) = N_0 \exp\left(-\frac{u}{1+r}\right) + \frac{N_0 R}{R+2}\left[\exp u - \exp\left(-\frac{u}{1+r}\right)\right]$$

(25)

The full solution follows by successive application of (23). There are two types of integrals that come up:

$$\exp\left(-\frac{m+1}{R+1}u\right) \int_0^u \frac{dv}{R+1} \left[m \exp v\right] \exp\left(\frac{m+1}{R+1}v\right)$$
$$= \frac{m}{R+2+m}\left[\exp u - \exp\left(-\frac{m+1}{R+1}u\right)\right] \quad (26)$$

$$\exp\left(-\frac{m+n+1}{R+1}u\right) \int_0^u \frac{dv}{R+1} (m+n)$$
$$\left[\exp\left(-\frac{mv}{R+1}\right)\left(1 - \exp\left(-\frac{v}{R+1}\right)\right)^n\right] \exp\left(\frac{m+n+1}{R+1}v\right)$$
$$= \frac{m+n}{n+1}\left[\exp\left(-\frac{mu}{R+1}\right)\left(1 - \exp\left(-\frac{u}{R+1}\right)\right)^{n+1}\right] \quad (27)$$

The coefficients that emerge from these integrations define the recursion relations for A_m and β_n^m:

$$A_{m+1} = \frac{m}{R+2+m}A_m$$

(28)

$$\beta_{n+1}^m = \frac{m+n}{n+1}\beta_n^m$$

(29)

The full solution to (2), taking into account the initial conditions, is given by:

$$F(m, t) = N_0 \phi^{-\frac{1}{1+R}}\left(1 - \phi^{-\frac{1}{1+R}}\right)^{m-1} + A_m\left(\phi - \phi^{-\frac{m}{1+r}}\right) - \sum_{i=1}^{m-1} A_i \beta_{m-i}^i \phi^{-\frac{i}{1+R}}\left(1 - \phi^{-\frac{1}{1+R}}\right)^{m-i}$$

$$A_m = \frac{RN_0}{R+2}\prod_{i=1}^{m-1}\frac{i}{R+2+i} = RN_0 \frac{\Gamma(m)\Gamma(R+2)}{\Gamma(R+2+m)}$$

$$\beta_n^m = \prod_{k=1}^{n}\frac{m+k-1}{k} = \frac{(m+n-1)!}{(m-1)!n!}$$

(30)

with the understanding that an empty product is unity, i.e. $\prod_{i=1}^{0} f(i) = 1$.

Note that the coefficients A_m and β_n^m do not depend on time, and furthermore β_n^m has no dependence on R or N_0. The product of coefficients $A_i \beta_{m-i}^i$ can be simplified:

$$A_i \beta_{m-i}^i = \frac{(m-1)!}{(m-i)! \prod_{j=1}^{i-1}(R+2+j)} \tag{31}$$

but it will be useful keep these these coefficients separate when considering the solution for more general initial conditions. Note that we use the standard definition for the gamma function $\Gamma(x)$; see Appendix H.
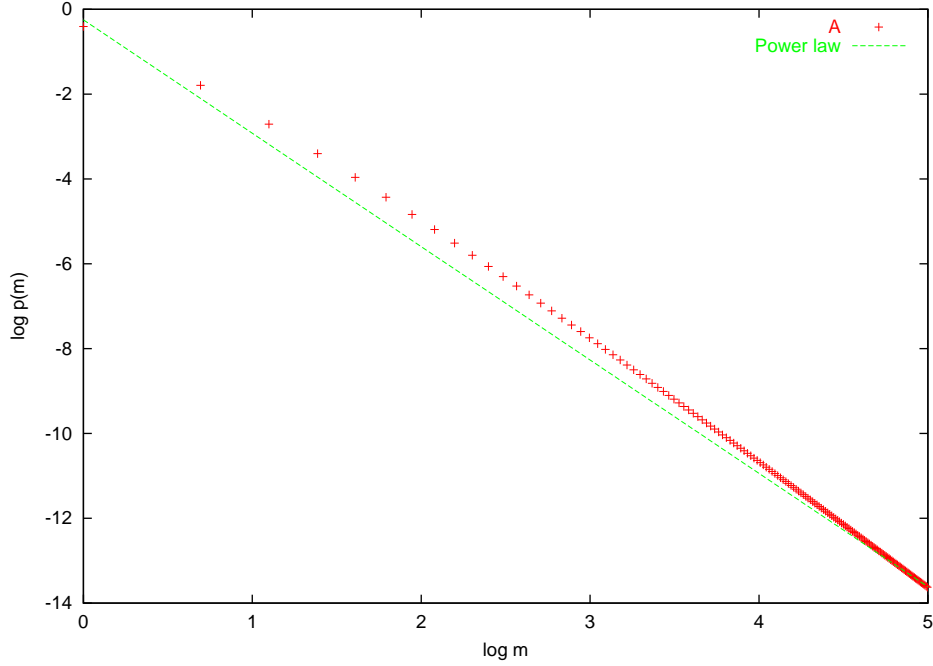

Figure 9: The normalized A_m coefficients (points) and a power-law fit (line), shown as a log-log plot as a function of size m, for $R = 1$.

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:

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

21

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 10 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 11 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.

C   Appendix: Arbitraty Initial Distribution

The solution for an arbitrary initial distribution: $N_{init}(m)$, requires solving (22) subject to different boundary conditions at $t = 0$; the terms proportional to A_m are the same, the term proportional to N_0 is replaced by the superposition of new terms describing the propagation of each bin of initial histogram:

$$F(m,t) = \sum_{i=1}^{\infty} N_{init}(i)\psi_i(m,t) + A_m(\phi - \phi^{-\frac{m}{1+r}}) - \sum_{i=1}^{m-1} A_i\psi_i(m,t)$$

$$\psi_i(m,t) = \begin{cases} 0 & \text{if } m < i \\ \beta_{m-i}^i \phi^{-\frac{i}{1+R}} \left(1 - \phi^{-\frac{1}{1+R}}\right)^{m-i} & \text{for } m \geq i \end{cases} \tag{35}$$

with the same definitions for A_m and β_n^m as before. These are derived by following by successive integration in the same way as was done in Appendix A.

The fact that $\psi_i(m,t) = 0$ for $m < i$ reflects the fact that there is no gene deletion; genes that start in bin i may either stay put or advance to bins corresponding to larger fold sizes, but will never populate bins of fold size less than i.
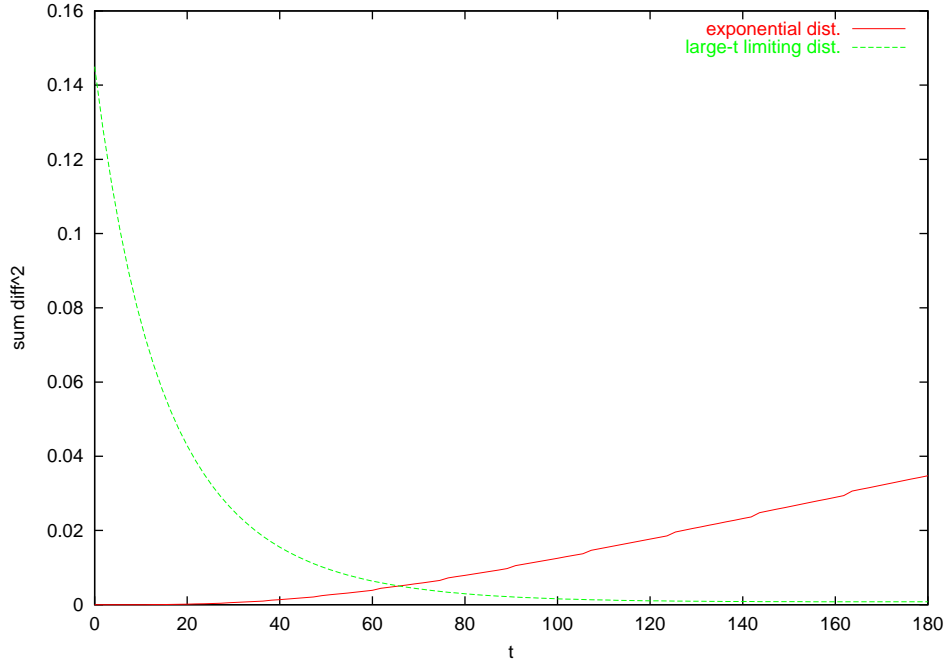
Figure 10: Crossover from exponential to large-time limiting distribution for $R = 1.0$ and $N_0 = 100$.

One important conclusion may be drawn from the full solution: all initial distributions ultimately lead to the the same limiting distribution determined by the A_m. Just as before, the dependence on the initial fold distribution $N_{init}(m)$ decays with time, leading to the same asymptotic distribution as was found for an initial distribution of N_0 folds of size 1 in (9). Of course, the details of how the crossover happens will depend on the particular form of $N_{init}(m)$.

D   Appendix: Solution to the Extended Model When $0 < Q < 1$ and $R = 0$

As one done in the solution for the minimal model, define $\phi(t)$:

$$\phi(t) = 1 + \frac{(1-Q)t}{N_0},\qquad(36)$$

and keep the association: $u = \log \phi(t)$. In terms of the time-like variable u, the fundamental evolution equations (10) now are:

$$(1-Q)\frac{\partial F(m,u)}{\partial u} = (m-1)F(m-1,t) - (1+Q)mF(m,t) + Q(m+1)F(m+1,u) \ \ (m > 1)$$

$$(1-Q)\frac{\partial F(1,u)}{\partial u} = -(1+Q)F(1,u) + QF(2,t)$$

$$(37)$$

Substituting the ansatz: $F(m,u) = f(u)g^{m-1}(u)$ into the equation for $m > 1$ leads
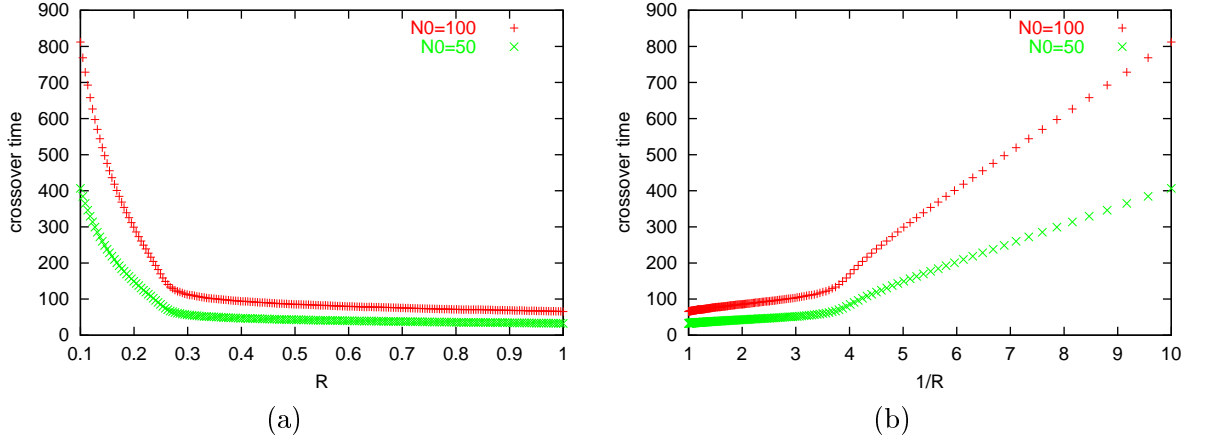
(a)   (b)

Figure 11: Crossover time for $N_0 = 100$ and $N_0 = 50$, plotted as a function of (a) R, and (b) $1/R$.

to the following relation:

$$(1-Q)\left[\frac{\partial \log f}{\partial u}g + (m-1)\frac{\partial g}{\partial u}\right] = (m-1) + (1+Q)mg + Q(m+1)g^2 \qquad (38)$$

Since neither $g(u)$ nor $f(u)$ depend on m, this identity can only be satisfied if:

$$(1-Q)\frac{\partial g}{\partial u} = 1 - (1+Q)g + Qg^2$$
$$(1-Q)\frac{\partial \log f}{\partial u} = -(1+Q) + 2Qg \qquad (39)$$

These equations can be solved by integration, together with the restriction that $f(t=0) = 1$ and $g(t=0) = 0$. It is easy to verify that the ansatz also works when $m = 1$.

$$F(m,t) = N_0 f(t) g^{m-1}(t)$$
$$f(t) = \phi^{-1}\left[\frac{1-Q}{1-Q\phi^{-1}}\right]^2 = \phi\left[\frac{N_0}{N_0+t}\right]^2$$
$$g(t) = \frac{1-\phi^{-1}}{1-Q\phi^{-1}} = \frac{t}{N_0+t} \qquad (40)$$
$$\phi(t) = 1 + \frac{(1-Q)t}{N_0}$$

In fact, it is easy to solve for $F(t)$ in this special case:

$$\frac{\partial F(t)}{\partial t} = -Q\frac{F(1,t)}{G(t)} = -Q\frac{f(t)}{\phi(t)} \qquad (41)$$

which can be integrated directly:

$$F(t) = N_0\frac{N_0 + (1-Q)t}{N_0+t} \qquad (42)$$

The large-time asymptotic limit for $F(t)$ is $(1-Q)N_0$ folds, which reflects the fact that some of the initial N_0 folds will ultimately be lost due to gene deletion.

24

Equation (42) leads to a simple relation between the number of folds and the number of genes:

$$F(t) = \frac{G(t)}{1 + t/N_0} \tag{43}$$

Although $F(t)$ and $G(t)$ both depend on Q, their ratio does not.

The solution (40) we have derived for $0 < Q < 1$ is also the solution for $Q = 1$, which means that gene deletion and duplication occur at the same rate. Equations (42) and (43) for the total number of folds $F(t)$ are still valid for $Q = 1$, but now the expected number of genes is constant: $G(t) = N_0$. Although we will not do so here, analytic solutions can be derived when deletion dominates duplication so the genome shrinks in size.

E   Appendix: Analytical Results for Higher Moments

Higher moments of the distribution, defined as $H_n(t) = \sum_m m^n F(m, t)$, for $n \geq 2$ in the extended model statisfy the following differential equation:

$$G(t)\frac{\partial H_n}{\partial t} = RG(t) + \sum_{m=1}^{\infty} F(m) \left[m(m+1)^n - (1+Q)m^{n+1} + Q(m-1)^n m \right] \tag{44}$$

In particular, the equations for the second and third moment are:

$$G(t)\frac{\partial H_2}{\partial t} = RG(t) + 2(1-Q)H_2(t) + (1+Q)G(t) \tag{45}$$

$$G(t)\frac{\partial H_3}{\partial t} = RG(t) + 3(1-Q)H_3(t) + 3(1+Q)3H_2(t) + (1-Q)G(t) \tag{46}$$

Higher moments depend on all lower moments except for the zeroth moment, the expected number of folds $F(t)$ This is fortuitous, since equation 11 for $F(t)$ could not be solved analytically due to its explicit dependence on the population of smallest folds: $F(1, t)$.

The solution for the second moment is given by:

$$H_2(t) = \begin{cases} N_0 \exp\left(\frac{2(1-Q)}{1+R-Q}u(t)\right) + N_0\frac{1+R+Q}{R-1+Q}\left[\exp u(t) - \exp\left(\frac{2(1-Q)}{1+R-Q}u(t)\right)\right] & R \neq 1-Q \\ N_0 \exp\left(u(t)\right)\left[1 + \frac{u(t)}{2R}\right] & R = 1-Q \end{cases} \tag{47}$$

where the variable $u(t)$ is related to the expected number of genes:

$$u(t) = \log\left[1 + \frac{(R+1-Q)t}{N_0}\right] \tag{48}$$

This result will be important in fitting actual genomic data to the models.

F   Appendix: Perturbation Theory Approximation for the Extended Model

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

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

25

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:

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

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)).

26

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

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

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 $i = 2$:

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

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 :

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

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 $m = 1$:

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

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:

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

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

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 12 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 $F(m, t)$.

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.
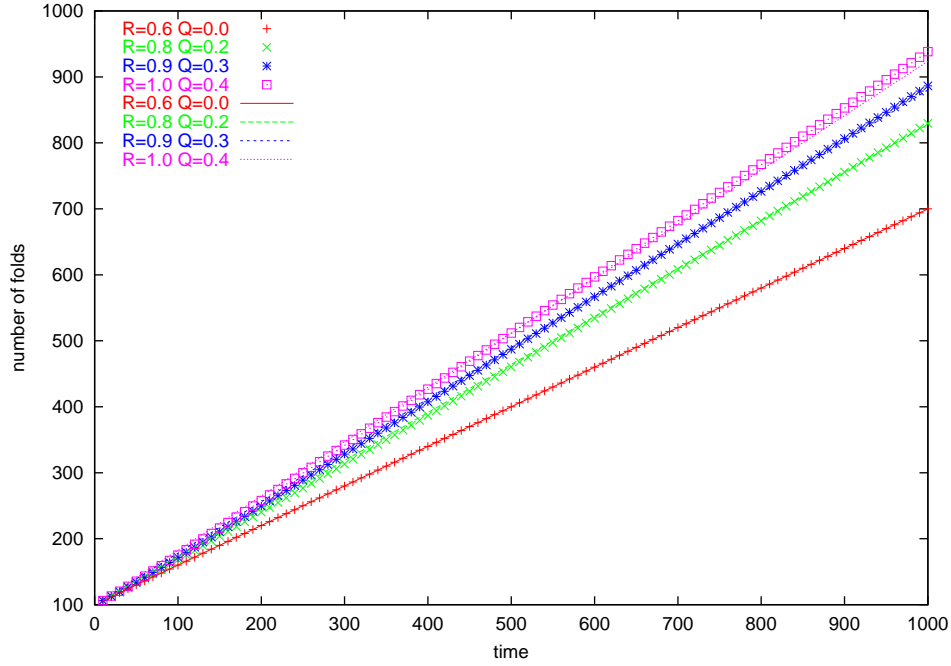
Figure 12: Analytic approximation for the total number of folds compared to numerical results of Figure 4.

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

29

(68):

$$\frac{\partial G_A(t)}{\partial t} = \frac{\partial}{\partial t} \sum_{m=1} m F_A(m,t) = R_A + \frac{G_A(t)}{G_A(t) + \gamma G_B(t)}$$

$$\frac{\partial G_B(t)}{\partial t} = \frac{\partial}{\partial t} \sum_{m=1} m F_B(m,t) = R_B + \gamma \frac{G_B(t)}{G_A(t) + \gamma G_B(t)} \qquad (69)$$

$$\frac{\partial G(t)}{\partial t} = \frac{\partial G_A(t)}{\partial t} + \frac{\partial G_B(t)}{\partial t} = R_A + R_B + 1$$

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

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

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:

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

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 $G_\gamma(t)$:

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

The full histogram is consequently a sum of exponential distributions:

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

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:

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

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:

$$C_m^A = \frac{R_A}{C_\gamma + 1} \prod_{i=1}^{m-1} \frac{i}{C_\gamma + i + 1}$$

$$C_m^B = \frac{R_B}{C_\gamma + \gamma} \prod_{i=1}^{m-1} \frac{i\gamma}{C_\gamma + \gamma(i+1)} \qquad (75)$$

30

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 (13). 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.
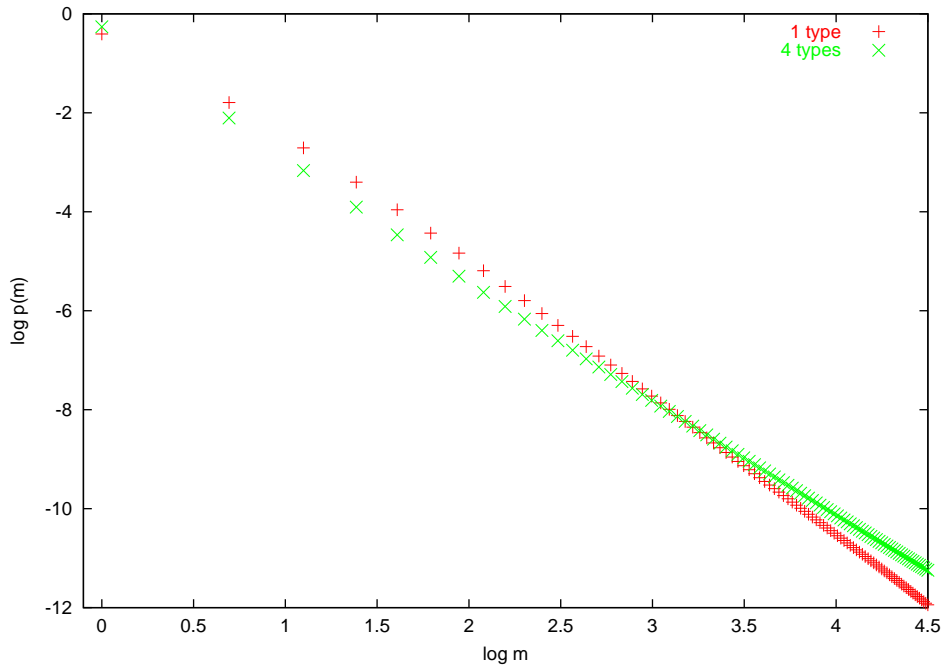

Figure 13: Large time limit for the fold probability distribution for the minimal model (one duplication type) and four duplication types: $\gamma_B = 4, \gamma_C = 8, \gamma_D = 16$. The total rate of new fold acquisition is the same for both genomes.

H Appendix: A Useful Normalization Identity

A series whose terms z_m, $m = 1, 2, \cdots$ are defined by a recursion relation:

$$z_m = \prod_{i=1}^{m-1} \frac{i}{\alpha + i} \qquad (77)$$

can be summed exactly as follows.

Rewrite z_m as:

$$z_m = \frac{\Gamma(m)\Gamma(\alpha + 1)}{\Gamma(\alpha + m)} \qquad (78)$$

with the usual definition for the gamma function:

$$\Gamma(z) = \int_0^\infty dt\, t^{z-1} e^{-t} \qquad (79)$$

The integral representation of the beta function $B(x, y)$ provides the key identity to carry out the sum:

$$B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x + y)} = \int_0^1 dt\, t^{x-1}(1 - t)^{y-1} \qquad (80)$$

Combining these relations leads to:

$$\sum_{m=1}^{\infty} z_m = \alpha \int_0^1 \sum_{m=1}^{\infty} t^{m-1}(1 - t)^{\alpha-1} \qquad (81)$$

$$= \alpha \int_0^1 (1 - t)^{\alpha-2} \qquad (82)$$

$$= \frac{\alpha}{\alpha - 1} \qquad (83)$$

$$(84)$$