Available online at www.sciencedirect.com
SCIENCE DIRECT®
Theoretical Population Biology 66 (2004) 93–100
Theoretical Population Biology
http://www.elsevier.com/locate/ytpbi

Subfunctionalization: How often does it occur? How long does it take?

Rachel Ward[a,1] and Richard Durrett[b,*,2]

[a] University of Texas, Austin, USA
[b] Department of Mathematics, Cornell University, 523 Malott Hall, Ithaca, NY 14850, USA

Received 3 December 2003

Available online 15 July 2004

Abstract

The mechanisms responsible for the preservation of duplicate genes have been debated for more than 70 years. Recently, Lynch and Force have proposed a new explanation: subfunctionalization—after duplication the two gene copies specialize to perform complementary functions. We investigate the probability that subfunctionalization occurs, the amount of time after duplication that it takes for the outcome to be resolved, and the relationship of these quantities to the population size and mutation rates.

Keywords: Gene duplication; Subfunctionalization

1. Introduction

Duplication of individual genes, chromosomal segments or even whole genomes, is a common occurrence in genome evolution and has historically been viewed as an important mechanism in the evolution of new gene functions (Ohno, 1970) or in providing protection against deleterious mutations (Clark, 1994; Nowak et al., 1997; Wagner, 1999). Despite the benefits of duplication, the mechanisms that lead to the preservation of these duplicate copies are not clear. There are three alternative outcomes in the evolution of duplicate genes: (i) one copy may be silenced by degenerative mutations; (ii) one copy may acquire a novel beneficial function; or (iii) a gene with two or more functions after duplication may result in a pair of genes specialized to perform complementary functions. For surveys, see Prince and Pickett (2002) and Walsh (2003).

Haldane (1933) argued that in the absence of fitness differences between the two copies, mutation would eventually inactivate one of the copies. A number of authors have investigated this model in which all individuals with at least one working copy are equally fit (Bailey et al., 1978; Kimura and King, 1979; Takahata and Maruyama, 1979; Watterson, 1983) and concluded that fixation takes a long time. However this observation cannot explain the high frequency of duplicates preserved in tetraploid fish, since there is no evidence that in these fishes the number of copies is polymorphic.

It is intuitively appealing that gene duplication allows the gene to experiment with mutations that would be deleterious to a single copy gene. However, theoretical work of Walsh (1995) has shown that preservation of duplicate copies due to positive selection is rare unless the ratio of favorable mutations to deleterious ones is not small and their effects are strongly beneficial. Recent studies of a number of fully sequenced genomes (see e.g., Lynch and Conrey, 2000 and Kondrashov et al., 2002) support the notion of a period of relaxed constraint by the observation that the ratio of nonsynonymous to synonymous substitutions is larger for young duplicates. Further evidence comes from a comparative study of 26 groups of orthologous genes from human, mouse,

*Corresponding author. Fax: +1-607-255-7149.
E-mail address: rtd1@cornell.edu (R. Durrett).

[1] R.W. is an undergraduate at University of Texas, Austin. This work was done while she participated in an NSF supported Research Experience for Undergraduates at Cornell in the summer of 2003.

[2] R.D. is partially supported by a grant from an NIGMS/NSF program for research at the interface between mathematics and biology.

0040-5809/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.tpb.2004.03.004

chicken, *Xenopus*, and zebrafish by Van de Peer et al. (2001) who found an increase in evolutionary rate in about half of the duplicated genes.

The third explanation introduced by Force et al. (1999) is that complementary degenerative mutations in the two copies lead to preservation of the duplicate copies. To explain this, consider a gene with two different functions controlled by different regulatory elements. If after duplication the first function is lost in the first gene and the second function lost in the second gene, then both genes are essential and will be preserved. This outcome is called *subfunctionalization*. Of course it is also possible for one copy to completely lose its functions while the other continues to perform both. We call this outcome *loss of function*.

In order to determine the relevance of subfunctionalization as an explanation of gene duplication, we need to understand what the model predicts. Lynch and Force (2000), see also Lynch et al. (2001), investigated the probability of subfunctionalization and the time until the outcome is determined in a Wright–Fisher model with constant population size N. In their model, each subfunction is lost with probability μ_r per generation and a gene loses all functions with probability μ_c. In this paper, we will take a closer look at these quantities for their model.

Lynch and Force only considered diploid individuals. However, as the reader will see, it is informative to compare the diploid case with the behavior in the haploid case. We will consider both unlinked and completely linked loci. The first case occurs when the duplicate copy ends up on a different chromosome, for example when a whole genome or chromosome is duplicated. The second case of complete linkage is a reasonable approximation for tandem duplication of genes, where the distance between copies is small. Finally, while Lynch and Force (2000) consider genes with more than two subfunctions we will restrict our attention to the case of two subfunctions.

Consider first the case in which the population size $N=1$. Given two copies the probability the first mutation results in loss of one subfunction in one copy is $4\mu_r/(4\mu_r+2\mu_c)$. Given this, the probability that the next mutation knocks out the complementary subfunction in the other gene is $\mu_r/(2\mu_r+\mu_c)$. The probability of subfunctionalization is thus

$$P_s = 2(\mu_r/(2\mu_r+\mu_c))^2.$$

If one notes that regulatory sequences are small compared to the size of genes and sets $\mu_r=0.1\mu_c$ then $P_s=0.01388$ is very small. However if we assume $\mu_r=\mu_c$, then $P_s=2/9=0.2222$. Lynch and Force (2000) performed simulations which show that the qualitative behavior of probability of subfunctionalization is the same when μ_r/μ_c is 3.0, 1.0, 0.3, or 0.1. Thus for simplicity, we only perform simulations for the case $\mu_r=\mu_c$. All of our theoretical calculations are for general μ_r and μ_c.

In Fig. 1 we have simulated the process with $\mu=10^{-4}$ and various values of N. As the top panel in Fig. 1 shows, $P_s=2/9$ provides a good estimate when N is small. To explain our choice of μ, we note that if a gene has 1000 nucleotides and we assume a per nucleotide per generation mutation rate of 10^{-8}, then $\mu=10^{-5}$. We have increased the mutation rate to $\mu=10^{-4}$ to speed up our simulations by a factor of 10. The answers are for large N determined by the value of $N\mu$, so in comparing with results of Lynch and Force (2000) one should compensate for our choice of mutation rate by dividing the population size by 10. That is, in our set up humans will roughly correspond to $N=1000$ and Drosophila to $N=100{,}000$.

To compute the mean time to resolution, ET, when $N=1$, we note that the mean time to the first mutation is $1/(4\mu_r+2\mu_c)$. If the first mutation is complete loss of function for one gene we are done. If not, an event of probability $2\mu_r/(2\mu_r+\mu_c)$, then the waiting time to the next event has mean $1/(2\mu_r+\mu_c)$. Adding the two parts

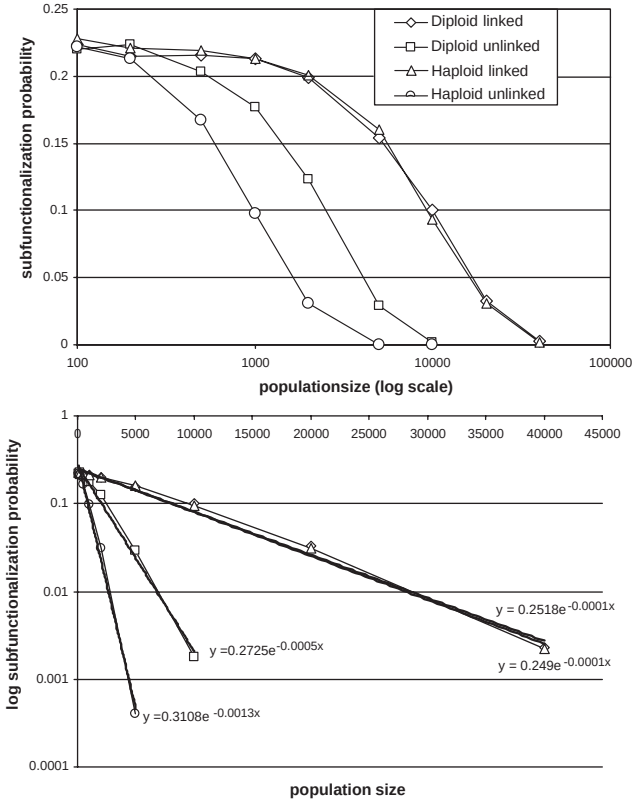

Fig. 1. Probability of subfunctionalization versus population size. In the top panel we have plotted probability versus log-population size as in Lynch and Force (2000). In the lower panel we have plotted log-probability versus population, which clearly shows the exponential decay.

gives

$$ET = \frac{6\mu_r + \mu_c}{2(2\mu_r + \mu_c)^2}.$$

When $\mu_r = \mu_c = \mu$, the expected time is $7/18\mu$. For $\mu = 10^{-4}$ this is approximately 3900 generations. As the top panel in Fig. 2 shows, this is accurate when N is small.

When $N > 1$ the last formula is an underestimate since it ignores the time for the mutation to become fixed in the population. Recalling that the average time until fixation for a mutation that is destined to fix is $2N$, suggests $7/18\mu + 2N$ as an improved approximation. This reasoning ignores various complexities associated with the process but, as Fig. 2 shows, it provides a reasonable approximation for the linked cases when N is 10,000 or less.

Lynch and Force (2000) plotted population size on a logarithmic scale. The results are clearer if one plots the population on an ordinary scale and in the first case plots the logarithm of P_s. See the lower panels of Figs. 1 and 2. In each case the data is almost a straight line indicating exponential decay of P_s as N increases in the first case and the linear growth of ET with population size in the second.

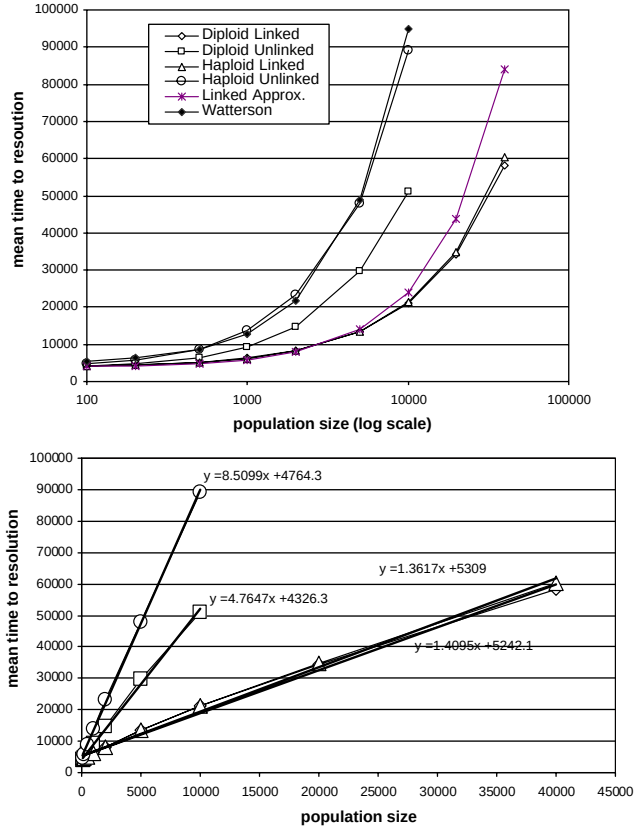

Fig. 2. Average time to resolution (subfunctionalization or loss of function) versus population size. In the top panel we have plotted average time versus log-population size as in Lynch and Force (2000). In the lower panel we have plotted average time versus population, which clearly shows the linear growth.

2. Linked loci

The main goal of this paper is to understand what happens in the subfunctionalization model when N is large. From Figs. 1 and 2 we see that the haploid and diploid linked cases show very similar behavior, so for simplicity we will consider the haploid linked model. Fig. 3 shows simulations with $\mu_r = \mu_c = 10^{-3}$ and $N = 2500$, and 40,000. Using four digit binary numbers to indicate the states of the two subfunctions in the two genes, there are nine possible states for viable individuals in the haploid linked model. To reduce the dimension we have color coded them as follows:

white	all working	1111,
yellow	3 out of 4 functions	1110, 1101, 1011, 0111,
green	subfunctionalization	1001, 0110,
red	loss of function	1100, 0011.

Fig. 3 shows that as the population size increases the model converges to the deterministic model in which offspring are produced with the expected frequencies. Writing x_3, x_2, x_1, and x_0 for the frequencies of white, yellow, green and red, and using $a = \mu_c$ and $a = \mu_r$ to simplify notation, the deterministic equations can be written as

$$\begin{aligned} x_3' &= x_3(1-2a-4b)/z, \\ x_2' &= 4bx_3 + x_2(1-2a-3b)/z, \\ x_1' &= bx_2 + x_1(1-2a-2b)/z, \\ x_0' &= 2ax_3 + (a+b)x_2 + x_0(1-a-2b)/z, \end{aligned}$$

where $z = x_3 + x_2(1-a-b) + x_1(1-2a-2b) + x_0(1-a-2b)$ is a normalization that makes the x' sum to 1.

To solve these equations it is convenient to consider instead

$$\begin{aligned} X_3' &= X_3(1-2a-4b), \\ X_2' &= 4bX_3 + X_2(1-2a-3b), \\ X_1' &= bX_2 + X_1(1-2a-2b), \\ X_0' &= 2aX_3 + (a+b)X_2 + X_0(1-a-2b). \end{aligned}$$

Since the original equations are linear except for the renormalization, it follows that if $X_i(n)$ and $x_i(n)$ are the values in generation n and $Z(n) = X_0(n) + X_1(n) + X_2(n) + X_3(n)$ then $x_i(n) = X_i(n)/Z(n)$.

The last set of equations is easy to solve. If we suppose that $X_3(0) = 1$ and the other $X_i(0) = 0$ then

$$X_3(n) = (1-2a-4b)^n.$$

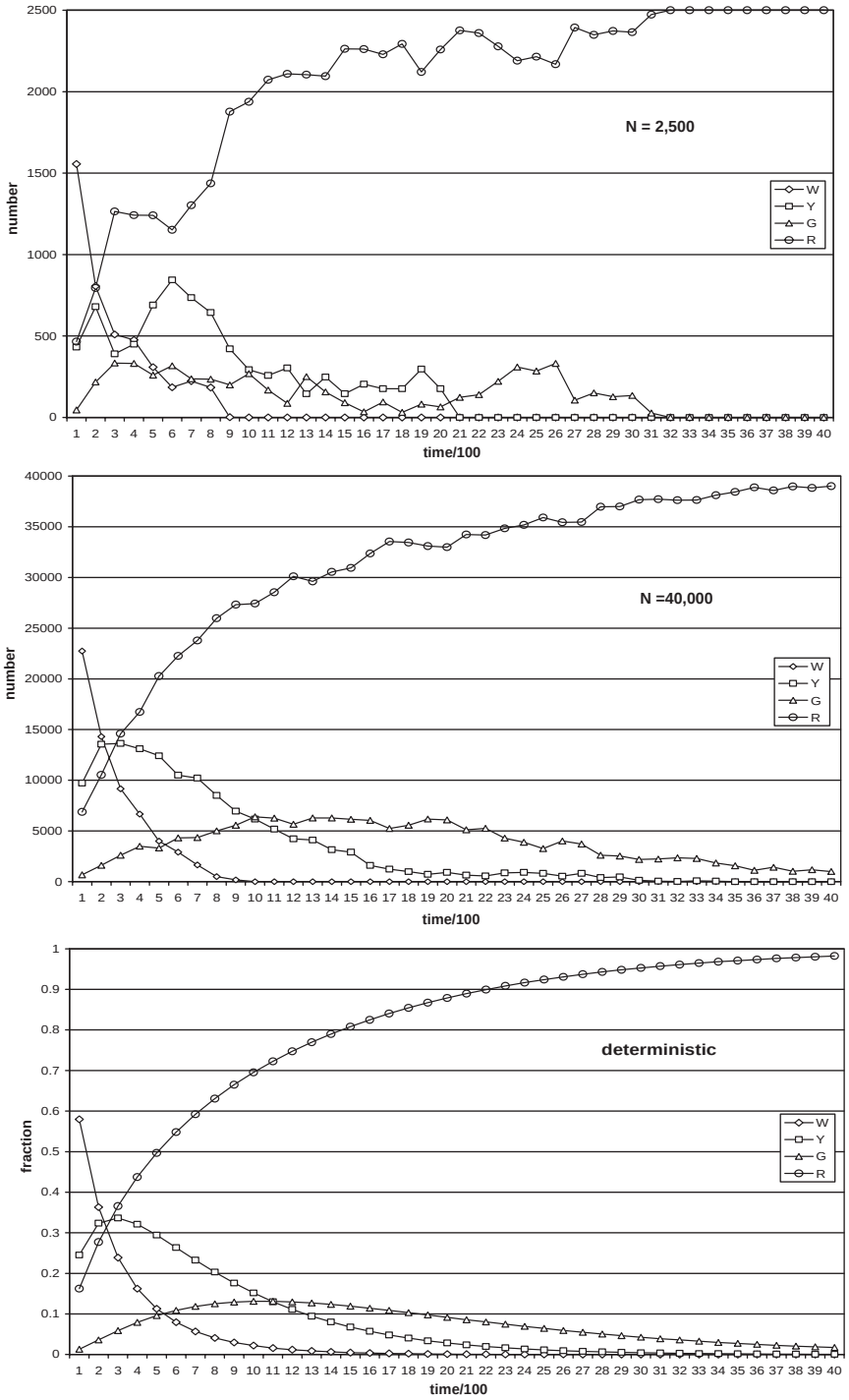

Fig. 3. Proportion of white, yellow, green, and red individuals (see Section 2 for definitions) in the haploid linked subfunctionalization model, showing convergence to the deterministic limit with increasing population size.

Using this in the second equation we have

$$X_2(n) = \sum_{m=0}^{n-1} 4b(1-2a-4b)^m(1-2a-3b)^{n-(m+1)}.$$

Bringing $4b(1-2a-3b)^{n-1}$ out front, we have

$$X_2(n) = 4b(1-2a-3b)^{n-1}\sum_{m=0}^{n-1}\left(1-\frac{b}{1-2a-3b}\right)^m.$$

Summing the geometric series, we have

$$X_2(n) = 4b\frac{(1-2a-3b)^n}{b}\left[1-\left(\frac{1-2a-4b}{1-2a-3b}\right)^n\right]$$
$$= 4[(1-2a-3b)^n-(1-2a-4b)^n].$$

Similar calculations that we have hidden away in the appendix give

$$X_1(n) = 2[(1-2a-2b)^n - 2(1-2a-3b)^n + (1-2a-4b)^n],$$

$$X_0(n) = \left(2+\frac{4b}{a+2b}\right)(1-a-2b)^n - 2(1-2a-2b)^n - 4(1-2a-3b)^n + 4\frac{a+b}{a+2b}(1-2a-4b)^n.$$

$X_0(n)$ decays more slowly than $X_1(n)$ so $x_0(n)\to 1$ as $n\to\infty$. Ignoring lower order terms and constants, in a population of size N we will have fixation when there are only a few 1s remaining and random fluctuations takeover, i.e., when

$$\frac{1}{N}\approx\frac{X_1(n)}{X_0(n)}\approx\left(1-\frac{a}{1-a-2b}\right)^n.$$

Approximating $1-a-2b\approx 1$, it follows that $n\approx(\log N)/a$. To see that it is permissible to ignore the lower order terms we note that the last equation implies

$$\frac{(1-2a-2b)^n}{(1-a-2b)^n}\approx 1/N, \quad \text{and} \quad \frac{(1-2a-3b)^n}{(1-2a-2b)^n}\approx 1/N^{b/a}.$$

The calculations above show that the deterministic system always ends with loss of function. In a population of size N, random fluctuations away from the deterministic trajectory are of order $1/\sqrt{N}$. The reasoning here is the same as the calculations. Kimura and King (1979) used to derive the diffusion approximation for the double null recessive model. We can thus regard the Lynch and Force model as a random perturbation of a dynamical system in the sense of Freidlin and Wentzell (1998). Their results, see e.g., Theorem 2.1 in Chapter 3, imply that for any fixed $\delta>0$ the probability of moving more than δ away from the deterministic trajectory is $\leqslant C\exp(-c(\delta)N)$ so in the linked cases the probability of subfunctionalization decay exponentially fast in N. Results in the next section will show that in the unlinked case the deterministic dynamics stay away from the subfunctionalization outcomes, so the exponential decay holds in that case as well.

3. Unlinked loci

In the linked cases when $N\mu$ is large, the system is almost deterministic and follows a path that leads to loss of function. The unlinked cases are more interesting. Let 3 = 11, 2 = 10, 1 = 01, and 0 = 00 denote the four possible states of a gene copy and let x_i and y_i denote the frequencies of states i and j at the first and second copy. The black diamonds in Fig. 4 give the values of (x_3,y_3) and of (x_3,x_2+x_1) at various time in five simulations of the diploid unlinked model with $N=10{,}000$ and $\mu=10^{-3}$.

To explain the curve of gray squares we will consider the deterministic dynamics. Writing $a=\mu_c$ and $b=\mu_r$ to simplify notation, the frequencies of gene 1 after the mutation step in either the haploid or the diploid model are

$$\xi_3 = x_3(1-a-2b),$$
$$\xi_2 = x_3 b + x_2(1-a-b),$$
$$\xi_1 = x_3 b + x_1(1-a-b),$$
$$\xi_0 = x_3 a + (x_1+x_2)(a+b) + x_0.$$

For the second locus there are similar formulas that describe the frequencies η_i after the mutation step.

In the haploid case the fraction of viable offspring is

$$w = \xi_3+\eta_3-\xi_3\eta_3+\xi_2\eta_1+\xi_1\eta_2$$

so the frequencies at the first copy in the next generation are

$$x_3 = \xi_3/w,$$
$$x_2 = \xi_2(\eta_3+\eta_1)/w,$$
$$x_1 = \xi_1(\eta_3+\eta_2)/w,$$
$$x_0 = \xi_0\eta_3/w$$

and the equations for the second copy are similar. Since the x_i and the y_i each sum to one there are six variables to solve for. It seems that we have six equations to determine the equilibria, but the conditions that x_3 and y_3 do not change in time lead to the same condition $w=1-a-2b$.

Based on the last observation, one might guess that there is a one-dimensional curve of possible equilibria. The first step in verifying this is to note that as $t\to\infty$ then $x_2(t)-x_1(t)\to 0$ and $y_2(t)-y_1(t)\to 0$. This is obvious from numerical computations and can be confirmed by algebraic manipulation of the equations for how x_2-x_1 and y_2-y_1 change in time (details not shown).

Once we know that in equilibria $x_2=x_1=x$ and $y_2=y_1=y$, the equations become somewhat simpler to solve. Letting $\alpha=1-a-2b$ and $\beta=1-a-b$ to simplify things the equations for x_3, x_2, and y_2 to be constant in time become

$$x_3+y_3-x_3y_3+2xy=\alpha,$$
$$\alpha x = (bx_3+\beta x)\beta(y_3+y),$$
$$\alpha y = (by_3+\beta y)\beta(x_3+x).$$

Let e_i be the ith equation. By considering e_2-e_3 and $\beta^2 e_1-e_2-e_3$ we get two equations which for fixed values of x_3 and y_3 are linear in x and y. Solving gives that x and y are linear functions of x_3 and y_3. Plugging

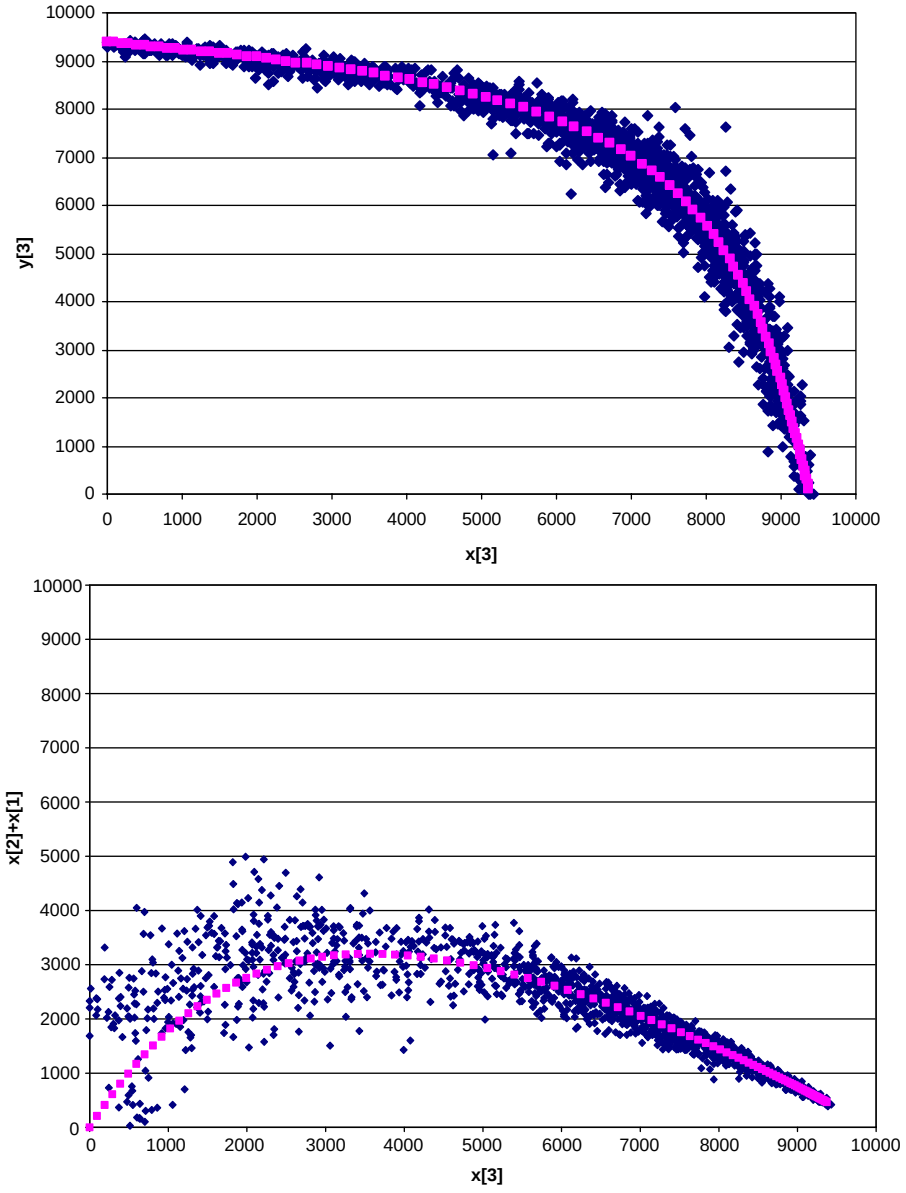

Fig. 4. Let 3 = 11, 2 = 10, 1 = 01, 0 = 00 denote the four possible states of the gene locus, and x and y the frequencies for the two unlinked loci. The graphs give frequencies of the indicated quantities at various times in five simulations of the diploid unlinked model.

the result into the first equation gives a relation between x_3 and y_3 in which the highest power of either variable is 2, so we can use the quadratic equation to solve for y_3 as a function of x_3.

The resulting expressions are very messy, but our analysis has established that for the haploid unlinked case there is a one-dimensional set of equilibria parameterized by $x_3 \in [0,1]$. Turning to the diploid unlinked case, the changes due to mutation are the same, but the greater number of gene copies makes the computations much more difficult. Let $u_1 = \xi_1 + \xi_0$ and $u_2 = \xi_2 + \xi_0$ be the probabilities that after mutation functions 1 and 2 do not exist in the first copy and $v_1 = \eta_1 + \eta_0$ and $v_2 = \eta_2 + \eta_0$ be the corresponding quantities for the second copy. Let F_i be the event that function i is present in at least one of the four copies. The total fraction of viable offspring,

$$\begin{aligned} w &= P(F_1 \cap F_2) = 1 - P(F_1^c \cup F_2^c) \\ &= 1 - P(F_1^c) - P(F_2^c) + P(F_1^c \cap F_2^c) \\ &= 1 - u_1^2 v_1^2 - u_2^2 v_2^2 + (\xi_0 \eta_0)^2. \end{aligned}$$

The frequencies at the first copy in the next generation are then

$$\begin{aligned} x_3 &= \xi_3 / w, \\ x_2 &= \xi_2 (1 - u_2 v_2^2)/w, \\ x_1 &= \xi_1 (1 - u_1 v_1^2)/w, \\ x_0 &= \xi_0 (1 - u_1 v_1^2 - u_2 v_2^2 + \xi_0 \eta_0^2)/w. \end{aligned}$$

Again our system is six-dimensional but there are only five equations since the ones for x_3 and y_3 to remain constant both reduce to $w = 1 - a - 2b$. This equation may look familiar, but is different from the previous one. In the current notation, the fraction of viable offspring in the haploid case is

$$w = 1 - u_1 v_1 - u_2 v_2 + \xi_0 \eta_0.$$

Thus in contrast to the linked case, the quantitative behavior of the haploid and diploid models is different.

We expect a one parameter family of solutions in the diploid unlinked case, but we are not able to algebraically solve the resulting equations. Fig. 4 however provides a numerical verification. The squares plotted there are the result of running the deterministic dynamics starting from a number of different starting points and all the values lie on a single curve.

The phenomenon we have observed in the unlinked cases is reminiscent of Watterson's (1983) computation for the double recessive null model of gene duplication, which is just the special case of the Lynch and Force model in which each gene has only one function. In Watterson's case there are only two variables, $x =$ the fraction of individuals in which copy 1 is functioning and $y =$ the fraction for copy 2.

In Watterson's case the state of the system moves quickly to within $1/N^{1/2}$ of the curve $xy = \mu^{1/2}$. By decomposing the motion into a component along the curve and one perpendicular to it, Watterson was able to approximate the time to loss of one copy as

$$N(\log(2N) - \psi(\theta/2)),$$

where $\theta = 4N\mu$ is the rescaled mutation rate, ψ is the digamma function

$$\psi(x) = -\gamma + \sum_{i=0}^{\infty} \frac{1}{i+1} - \frac{1}{i+x}$$

and $\gamma \approx 0.57721$ is Euler's constant.

Given that the qualitative behavior of the subfunctionalization model is similar to that of Watterson's process, it is not unreasonable to hope (see Fig. 5 in Lynch and Force, 2000) that $N(\log(2N) - \psi(\theta/2))$ will give a good approximation to the time to resolution. However, this is just a hope since there is no reason that Watterson's proof for the one function case, which depends on the specific form of the equilibrium curve in that case, will generalize to the case of two subfunctions. Consulting Fig. 3 we see that it gives a reasonable approximation to the haploid unlinked case but not to the values for the diploid unlinked case, in contrast to what is shown in Fig. 5 of Lynch and Force (2000).

4. Summary

We have investigated properties of the subfunctionalization model of Lynch and Force (2000). When the population size times the mutation rate, $N\mu$, is small the probability of subfunctionalization and the time until resolution are well approximated by the values when $N = 1$. When $N\mu$ gets large the subfunctionalization probability decays exponentially fast to 0.

In the linked cases, the frequencies of chromosomes of various types are quantitatively similar in the haploid and diploid cases. Simulations suggest that the time to resolution grows linearly with $c = N\mu$ but analytical results show that for fixed mutation rate the time grows like $\log N$ when N is large.

In the unlinked cases, chromosome frequencies are different in the haploid and diploid cases but the models are qualitatively similar. When $N\mu$ is large, the observed values of the states of the two copies stay close to a one dimensional subset of the six-dimensional space of probabilities. This phenomenon is similar to what Watterson (1983) observed for the double recessive null model. His formula works well to predict the time to resolution in the haploid unlinked case but not in the diploid case.

Appendix

Using the formula for $X_2(n)$ in the equation for $X_1(n)$ gives

$$\begin{aligned} X_1(n) &= \sum_{m=0}^{n-1} 4b[(1-2a-3b)^m \\ &\quad - (1-2a-4b)^m](1-2a-2b)^{n-m-1} \\ &= 4b(1-2a-2b)^{n-1} \sum_{m=0}^{n-1} \left(1 - \frac{b}{1-2a-2b}\right)^m \\ &\quad - \left(1 - \frac{2b}{1-2a-2b}\right)^m. \end{aligned}$$

The sum is equal to

$$\begin{aligned} &\frac{1-2a-2b}{b}\left[1 - \left(\frac{1-2a-3b}{1-2a-2b}\right)^n\right] \\ &\quad - \frac{1-2a-2b}{2b}\left[1 - \left(\frac{1-2a-4b}{1-2a-2b}\right)^n\right]. \end{aligned}$$

Multiplying and dividing by 2 in the first term we can combine the two terms. Inserting the result into the previous equation gives

$$\begin{aligned} X_1(n) = 2[&(1-2a-2b)^n - 2(1-2a-3b)^n \\ &+ (1-2a-4b)^n]. \end{aligned}$$

Using the formula for $X_2(n)$ in the equation for $X_0(n)$ gives

$$X_0(n) = \sum_{m=0}^{n-1} 2a(1-2a-2b)^m(1-a-2b)^{n-m-1} + \sum_{m=0}^{n-1}(a+b)4[(1-2a-3b)^m - (1-2a-4b)^m](1-a-2b)^{n-m-1}.$$

Doing the sums gives

$$X_0(n) = 2a(1-a-2b)^{n-1}\frac{1-a-2b}{a} \times\left[1-\left(\frac{1-2a-2b}{1-a-2b}\right)^n\right] + 4(a+b)(1-a-2b)^{n-1} \times\left\{\frac{1-a-2b}{a+b}\left[1-\left(\frac{1-2a-3b}{1-2a-2b}\right)^n\right] - \frac{1-a-2b}{a+2b} \times\left[1-\left(\frac{1-2a-4b}{1-2a-2b}\right)^n\right]\right\}.$$

After a little algebra this reduces to

$$X_0(n) = \left(2+\frac{4b}{a+2b}\right)(1-a-2b)^n - 2(1-2a-2b)^n - 4(1-2a-3b)^n + 4\frac{a+b}{a+2b}(1-2a-4b)^n.$$

References

Bailey, G.S., Poulter, R.T.M., Stockwell, P.A., 1978. Gene duplication in tetraploid fish: model for gene silencing at unlinked duplicated loci. Proc. Natl. Acad. Sci. USA 75, 5575–5579.

Clark, A., 1994. Invasion and maintenance of a gene duplication. Proc. Natl. Acad. Sci. USA 91, 2950–2954.

Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y., Postlethwait, J., 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151, 1531–1545.

Freidlin, M.I., Wentzell, A.D., 1988. Random Perturbations of Dynamical Systems. Springer, New York.

Haldane, J.B.S., 1933. The part played by recurrent mutations in evolution. Amer. Nat. 67, 5–19.

Kimura, M., King, J.L., 1979. Fixation of deleterious allele at one of two duplicate loci by mutation pressure and random drift. Proc. Natl. Acad. Sci. USA 76, 2858–2861.

Kondrashov, F.A., Rogozin, I.B., Wolf, Y.I., Koonin, E.V., 2002. Selection in the evolution of gene duplications. Genome Biol. 3(2), research0008.1-9.

Lynch, M., Conrey, J.S., 2000. The evolutionary fate and consequences of duplicated genes. Science 290, 1151–1155.

Lynch, M., Force, A., 2000. The probability of preservation of a duplicate gene by subfunctionalization. Genetics 154, 459–473.

Lynch, M., O'Hely, M., Walsh, B., Force, A., 2001. The probability of the preservation of a newly arisen gene duplicate. Genetics 159, 1789–1804.

Nowak, M.A., Boerlijst, M.C., Cooke, J., Maynard Smith, J., 1997. Evolution of genetic redundancy. Nature 388, 167–171.

Ohno, S., 1970. Evolution by Gene Duplication. Springer, Berlin.

Prince, V.E., Pickett, F.B., 2002. Splitting pairs: the diverging fates of duplicated genes. Nat. Rev. Genet. 3, 827–837.

Takahata, N., Maruyama, T., 1979. Polymorphism and loss of duplicate gene expression: a theoretical study with application to tetraploid fish. Proc. Natl. Acad. Sci. USA 76, 4521–4525.

Van de Peer, Y., Taylor, J.S., Braasch, I., Meyer, A., 2001. The ghost of selection past: rates of evolution and functional divergence of anciently duplicated genes. J. Mol. Evol. 53, 436–446.

Wagner, A., 1999. Redundant gene functions and natural selection. J. Evol. Biol. 12, 1–16.

Walsh, J.B., 1995. How often do duplicated genes evolve new function? Genetics 139, 421–428.

Walsh, B., 2003. Population-genetic models of the fates of duplicated genes. Genetica 118, 279–294.

Watterson, G.A., 1983. On the time for gene silencing at duplicate loci. Genetics 105, 745–766.