Waiting for two mutations: with applications to regulatory sequence evolution and the limits of Darwinian evolution

Rick Durrett* and Deena Schmidt†

*Department of Mathematics and †Center for Applied Mathematics,

Cornell University, Ithaca, New York 14853

August 21, 2008

Running Head: Waiting for two mutations

Key words: regulatory sequence evolution, Moran model, Drosophila, humans

Corresponding Author:

Rick Durrett

523 Malott Hall

Department of Mathematics, Cornell University

Ithaca, NY 14853

Phone: (607) 255-8282

Fax: (607) 255-7149

Email: rtd1@cornell.edu

ABSTRACT

Results of Nowak and collaborators concerning the onset of cancer due to the inactivation of tumor suppresor genes give the distribution of the time until some individual in a population has experienced two prespecified mutations, and the time until this mutant phenotype becomes fixed in the population. In this article we apply these results to obtain insights into regulatory sequence evolution in Drosophila and humans. In particular, we examine the waiting time for a pair of mutations, the first of which inactivates an existing transcription factor binding site and the second which creates a new one. Consistent with recent experimental observations for Drosophila, we find that a few million years is sufficient, but for humans with a much smaller effective population size, this type of change would take more than 100 million years. In addition, we use these results to expose flaws in some of Michael Behe's arguments concerning mathematical limits to Darwinian evolution.

INTRODUCTION

There is a growing body of experimental evidence that in Drosophila, significant changes in gene regulation can occur in a short amount of time, compared to divergence time between species. LUDWIG *et al.* (1998, 2000, 2005) studied the evolution of the *even-skipped* stripe 2 enhancer in four Drosophila species (*D. melanogaster*, *D. yakuba*, *D. erecta*, and *D. pseudoobscura*). While expression is strongly conserved, they found many substitutions in the binding sites for bicoid, hunchback, Kruppel, and giant, as well as large differences in the overall size of the enhancer region. In addition, they uncovered several binding sites that have been gained and lost among these four species: a lineage-specific addition of the bicoid-3 binding site in *D. melanogaster* that is absent in the other species, a lineage-specific loss of the hunchback-1 site in *D. yakuba*, and the presence of an extra Kruppel site in *D. pseudoobscura* relative to *D. melanogaster*. These differences are nicely summarized in Figure 2B of LUDWIG *et al.* (2005).

In a simulation study, STONE and WRAY (2001) estimated the rate of *de novo* generation of regulatory sequences from a random genetic background. They found that for a given 6 nucleotide sequence, the time until it arose in a 2 kb region in some individual was about 5,950 years for humans and 24 years for Drosophila. However, as MACARTHUR and BROOKFIELD (2004) have already pointed out, there is a serious problem with Stone and Wray's computation. They assumed individuals in the population evolve independently, while in reality there are significant correlations due to their common ancestors.

Motivated by this simulation study, DURRETT and SCHMIDT (2007) have recently given a mathematical analysis for regulatory sequence evolution in humans, correcting the calculation mentioned above. They assumed an effective population size of 10,000 and a per nucleotide mutation rate of $\mu = 10^{-8}$. In this situation, the expected number of segregating

sites in a 1 kb sequence is $1000(4N_e\mu) = 0.4$ so it makes sense to talk about a population consensus sequence. The authors defined this as the nucleotide at the site if there is no variability in the population and if the site is variable, the most frequent nucleotide at that site in the population. Using a generation time of 25 years, they found that in a 1 kb region, the average waiting time for words of length 6 was 100,000 years. For words of length 8, they found that the average waiting time was 375,000 years when there was a 7 out of 8 letter match to the target word in the population consensus sequence (an event of probability roughly 5/16) and 650 million years when there was not.

Fortunately, in biological reality, the match of a regulatory protein to the target sequence does not have to be exact for binding to occur. Biological reality is complicated, with the acceptable sequences for binding described by position weight matrices that indicate the flexibility at different points in the sequence. To simplify, we will assume that binding will occur to any 8 letter word that has seven letters in common with the target word. If we do this, then the mean waiting time reduces to about 60,000 years.

To explain the intuition behind the last result, consider the case of 8 letter words. If all nucleotides are equally likely and independent, then using the binomial distribution we see that a 6 out of 8 match to a given 8 letter target word has probability $63/16384 \approx 0.00385$, so in a region of 1000 nucleotides, we expect to find 3.85 such approximate matches in the population consensus sequence. Simple calculations then show that the waiting time to improve one of these 6 out of 8 matches to 7 out of 8 has a mean of 60,000 years. This shows that new regulatory sequences can come from small modifications of existing sequence.

Extending our previous work on the *de novo* generation of binding sites, this article will consider the possibility that in a short amount of time, two changes will occur, the first of which inactivates an existing binding site, and the second which creates a new one. This

problem was studied earlier by CARTER and WAGNER (2002). In the next section, we present the model and then a simpler theoretical analysis based on work of KOMAROVA *et al.* (2003) and IWASA *et al.* (2004, 2005), who studied the onset of cancer due to the inactivation of tumor suppressor genes. Lastly, we compare the theory with simulations and experimental results.

THE MODEL

Consider a population of $2N$ haploid individuals. The reader should think of this as the chromosomes of N diploid individuals evolving under the assumptions of random union of gametes and additive fitness. However, since we will use the continuous time Moran model, it is simpler and clearer to state our results for haploid individuals.

We start with a homogeneous population of wild type individuals. We have two sets of possible mutant genotypes A and B. Wild type individuals mutate to type A at rate u_1 and type A individuals mutate to type B at rate u_2. We assume there is no back mutation. We think of the A mutation as damaging an existing transcription factor binding site and the B mutation as creating a second new binding site at a different location within the regulatory region. We assign relative fitnesses 1, r, and s to wild type, A mutant, and B mutant individuals respectively. See Figure 1 for a diagram of our model. We have used the word damage above to indicate that the mutation may only reduce the binding efficiency, not destroy the binding site. However, even if it does, the mutation need not be lethal. In most cases the B mutation will occur when the number of A mutants is a small fraction of the population, so most individuals with the A mutation will also carry a working copy of the binding site.

We could also assume that the mutations occur in the other order: B first then A. This is also a two stage process which falls into the general framwork of our analysis below under

the appropriate fitness assumptions. The problems in population genetics to be solved are: how long do we have to wait for (i) a type B mutation to arise in some individual? or (ii) for the B mutant to become fixed in the population? These problems were investigated by KOMAROVA *et al.* (2003) and IWASA *et al.* (2004, 2005) for tumor suppressor genes, whose inactivation can lead to cancer, with A being the inactivation of one copy of the gene and B the inactivation of the other. A nice account of these results can be found in Section 12.4 of NOWAK's (2006) book on evolutionary dynamics. Here, we will apply these results to estimate the waiting time for a switch between two transcription factor binding sites, as defined in the statement of our problem above.

First we need to describe the population genetics model we are considering. Rather than use the discrete time Wright-Fisher model, we use the continuous time Moran model. We prefer the Moran process because it is a birth and death chain which means that the number of type A individuals increases or decreases by one on each event. Biologically the Moran model corresponds to a population with overlapping generations, and in the case of tumor suppresor genes is appropriate for a collection of cells in an organ which is being maintained at a constant size.

As the reader will see from the definition, the Moran process as a genetic model treats N diploids as $2N$ haploids and replaces one chromosome at a time. In this context it is common to invoke random union of gametes and assume fitnesses are additive, but that is not necessary. Since heterozygotes are rare, the fitness of an A mutant is its fitness in the heterozygous state. Supposing that the relative fitnesses have been normalized to have maximum value 1, the dynamics may be described as follows:

- Each individual is subject to possible replacement at rate 1.
- A copy is made of an individual chosen at random from the population.

- Mutation changes the copy from wild type to A with probability u_1 and from A to B with probability u_2.
- If the relative fitness of the proposed new individual is 1-q after mutation (where q is the selection coefficient), then the replacement occurs with probability 1-q. Otherwise, nothing happens.

For more on this model, see Section 3.4 of EWENS (2004).

THEORETICAL RESULTS

Neutral case: Returning to the problem, we first consider the case in which the fitness of the A mutant $r = 1$ and the population is of intermediate size as compared to the mutation rates. That is, one in which the population size and mutation rates satisfy

$$1/\sqrt{u_2} \ll 2N \ll 1/u_1 \tag{1}$$

where for any numbers a and b, $a \ll b$ is read "a is much less than b" and means a/b is small.

Theorem 1. *If $2N \ll 1/u_1$ and $2N \gg 1/\sqrt{u_2}$, the probability $P(t)$ that a B mutation has occurred in some member of the population by time t*

$$P(t) \approx 1 - \exp(-2Nu_1\sqrt{u_2}t) \tag{2}$$

where $\approx$ is read "approximately." If B mutants become fixed with probability β then the result for the fixation time is obtained by replacing u_2 by βu_2 in (2).

In words, the waiting time τ_B for the first B mutation is roughly exponential with mean $1/(2Nu_1\sqrt{u_2})$ while the waiting time T_B for B to become fixed is roughly exponential with

mean $1/(2Nu_1\sqrt{u_2\beta})$. Hence, if $\beta < 1$, this increases the waiting time by a factor of $1/\sqrt{\beta}$ rather than the $1/\beta$ that one might naively expect. The last conclusion in the theorem should be intuitive since successful mutations (i.e., those that go to fixation) occur at rate βu_2. In each of the next three theorems, the results for the fixation time can be obtained by replacing u_2 by βu_2 in the waiting time result.

Sketch of proof. The mathematical proof of this result involves some technical complications, but the underlying ideas are simple. Here and in what follows, readers not interested in the underlying theory can skip the proof sketches. A simple calculation, see section 2 of IWASA *et al.* (2005), shows that the probability a type A mutant will give rise to a type B mutant before its family dies out is asymptotically $\sqrt{u_2}$. Since type A mutations arise at rate $2Nu_1$, the time σ_B until an A mutant arises that will have a descendant of type B is exponential with mean $1/(2Nu_1\sqrt{u_2})$. The amount of time after σ_B it takes for the B mutant to appear, $\tau_B - \sigma_B$, is of order $1/\sqrt{u_2}$. Since $2Nu_1 \ll 1$, the difference $\tau_B - \sigma_B$ is much smaller than σ_B and the result holds for τ_B as well.

To give some intuition about how the B mutation arises, we note that in the neutral case, $r = 1$, the number of copies of the A allele increases and decreases with equal probability, so if we ignore the transitions that don't change the number of mutant alleles, the result is an unbiased random walk. Since such a walk represents the winnings of a gambler playing a fair game, the probability that the number will rise to $k = 1/\sqrt{u_2}$ before hitting 0 is $1/k$. When this occurs the central limit theorem of probability theory implies that the number of steps required to return to 0 is of order $k^2 = 1/u_2$, since this requires the random walk to move by k, and by the central limit theorem this will take time of order k^2. Since B mutants have probability u_2, there is a reasonable chance of having a B mutation before the number of A mutants returns to 0. □

IWASA *et al.* (2004) call this *stochastic tunneling*, since the second mutant (type B) arises before the first one (type A) fixes. CARTER and WAGNER (2002) had earlier noticed this possibility but they did not end up with a very nice formula for the average fixation time, see their (2.2) and the formulas for the constants given in their appendix. The assumption $1/\sqrt{u_2} \ll 2N$ implies that throughout the scenario we have just described, the number of type A mutants is a small fraction of the population, so we can ignore the probability that the A mutants become fixed in the population. This means that in an intermediate sized population (1) with $r = 1$, B mutations arise primarily through stochastic tunneling.

In contrast to populations of intermediate size, populations that are small (compared to the mutation rates) have fixation of the type A mutation before the type B mutation arises.

Theorem 2. *If $2N \ll 1/u_1$ and $2N \ll 1/\sqrt{u_2}$ then the probability $P(t)$ that a B mutation has occurred in some member of the population by time t*

$$P(t) \approx 1 - \exp(-u_1 t) \tag{3}$$

Sketch of proof. To explain this, we note that the waiting times between A mutations are exponential with mean $1/(2Nu_1)$ and each one leads to fixation with probability $1/(2N)$, so the time we have to wait for the first A mutation that will go to fixation is exponential with mean $1/u_1$. The condition $2N \ll 1/\sqrt{u_2}$ implies that it is unlikely for the B mutation to appear before A reaches fixation. The average time required for an A mutation to reach fixation conditional on fixation, which is $2N$ by a result of KIMURA and OHTA (1969), and the average time required for the B mutation to appear after fixation of the A mutation, which is $1/(2Nu_2)$, are each short compared to $1/u_1$ and can be ignored. □

When $2N$ and $1/\sqrt{u_2}$ are about the same size, fixation of an A mutation and stochastic tunneling are both possible situations in which a type B mutation can arise, and the analysis

becomes very complicated. IWASA *et al.* (2005) obtained some partial results, see their equation (15). Recent work of Durrett, Schmidt, and Schweinsberg (2007) addresses this borderline case and gives the following result.

Theorem 3. *If* $2N \ll 1/u_1$ *and* $2N\sqrt{u_2} \to \sqrt{\gamma}$*, then the probability* $P(t)$ *that a* B *mutation has occurred in some member of the population by time* t*,* $P(t) \approx 1 - \exp(-\alpha(\gamma)u_1 t)$ *where*

$$\alpha(\gamma) = \sum_{k=1}^{\infty} \frac{\gamma^k}{(k-1)!(k-1)!} \Bigg/ \sum_{k=1}^{\infty} \frac{\gamma^k}{k!(k-1)!} > 1 \tag{4}$$

Hence, the mean waiting time in this case is $1/(\alpha(\gamma)u_1)$.

To summarize the first three theorems, if $2N \ll 1/u_1$ then the waiting time for the first B mutant to appear in the population, τ_B, is approximately exponential under the following conditions:

	assumption	$E\tau_B$
Theorem 1	$2N \gg 1/\sqrt{u_2}$	$1/2Nu_1\sqrt{u_2}$
Theorem 2	$2N \ll 1/\sqrt{u_2}$	$1/u_1$
Theorem 3	$2N\sqrt{u_2} \to \sqrt{\gamma}$	$1/(\alpha(\gamma)u_1)$

Deleterious A mutants: Suppose now that A mutants have fitness $r < 1$, and return to the case of intermediate population size defined by (1).

Theorem 4. *Suppose* $1/\sqrt{u_2} \ll 2N \ll 1/u_1$ *and that* $1 - r \approx \rho\sqrt{u_2}$*, where* ρ *is a constant that measures the strength of selection against type* A *mutants. The probability* $P(t)$ *that the* B *mutation has occurred in some member of the population by time* t

$$P(t) \approx 1 - \exp(-2Nu_1 R\sqrt{u_2}t) \quad \textit{where} \quad R = \frac{1}{2}\left(\sqrt{\rho^2+4} - \rho\right) \tag{5}$$

The proof of this is somewhat involved so we refer the reader to IWASA *et al.* (2005) for details. In words, the conclusion says that the waiting time in the non-neutral case is still

exponential, but the mean has been multiplied by $1/R$. Note that when $\rho = 0$, $R = 1$ which is the neutral case. Thus, A mutants are essentially neutral when $\rho \approx 0$ which is true when $1 - r \ll \sqrt{u_2}$. When ρ is large, $\sqrt{\rho^2 + 4} \approx \rho + 2/\rho$ (since $(\rho + 2/\rho)^2 = \rho^2 + 4 + 4/\rho^2$) and we have $1/R \approx \rho$. Therefore, as ρ increases the waiting time increases.

Kimura (1985) considered compensatory mutations which are related to the situation studied in Theorem 4. His model has four genotypes AB, $A'B$, AB', and $A'B'$, where A and B are wild type alleles with corresponding mutant alleles A' and B'. The single mutant genotypes $A'B$ and AB' have fitness $1 - s$ while AB and $A'B'$ have fitness 1. Assuming $s \gg v$, the mutation rate, he used diffusion theory to conclude that the average time for the fixation of the double mutant was, see his (16) and (17) and take $h = 0$,

$$4N_e \int_0^1 \exp(-S\eta^2/2)\eta^{-V} \int_0^\eta \exp(S\xi^2/2)\frac{\xi^{V-1}}{1-\xi}\, d\xi\, d\eta$$

where $S = 4N_e s$ and $V = 4N_e v$. Evaluating the expression above numerically he concluded that the fixation time was surprisingly short. Note that his result covers a different range of parameters since Theorem 4 supposes $4Nv \ll 1$. However, stochastic tunneling still occurs. Kimura shows that the frequency of single mutants remains small until the second mutation occurs.

SIMULATION RESULTS

The results in the previous section are theorems about the limit as $N \to \infty$, and their proofs are based on arguing that various complications can be ignored, so we will now turn to simulations to show that the approximations are good for even relatively small values of N. We will use a standard algorithm, described in the next paragraph, to simulate the continuous time Markov chain $X(t)$ which counts the number of A mutants in the population at time t. Readers not interested in the details of our simulation algorithm can skip the next

paragraph.

Let $T_0 = 0$ and for $m \geq 1$ let T_m be the time of the mth jump of $X(t)$. If $X(T_m) = 0$, we let t_{m+1} be exponential with mean $1/(2Nu_1)$, and set $T_{m+1} = T_m + t_{m+1}$ and $X(T_{m+1}) = 1$. If $X(T_m) = k$ with $1 \leq k < 2N$, then we let t_{m+1} be exponential with mean $1/(p_k + q_k + r_k)$ where p_k is the rate of jumps to $k+1$, q_k is the rate of jumps to $k-1$, and r_k is the rate an A mutant replaces an A mutant, as defined in the following table. Note that the second term in p_k accounts for new A mutants that enter the population.

$$\begin{array}{lll} k \to k+1 & \text{at rate} & p_k = \dfrac{k(2N-k)}{2N} + (2N-k)u_1 \\ k \to k-1 & \text{at rate} & q_k = \dfrac{k(2N-k)}{2N} \\ k \to k & \text{at rate} & r_k = \dfrac{k^2}{2N} \end{array}$$

We set $X(T_{m+1}) = k+1$ with probability $p_k/(p_k + q_k + r_k)$, $X(T_{m+1}) = k$ with probability $r_k/(p_k + q_k + r_k)$, and $X(T_{m+1}) = k-1$ with probability $q_k/(p_k + q_k + r_k)$. In the first two cases there is a probability u_2 of a B mutation. We stop the simulation the first time a B mutant appears or $X(T_m) = 2N$. If an A mutant goes to fixation, we add an exponential with mean $1/(2Nu_2)$ to the final time to simulate waiting for the B mutation to appear.

Let $n = 2N$. Since our aim is to show that the theoretical predictions work well even for small values of n, we will, in most cases, consider the values $n = 1,000$ and $n = 10,000$. Table 1 gives the seven simulation scenarios we study. Table 2 compares the predicted mean time from Theorem 1 with the average time found in 10,000 replications of each simulation. In making predictions for the examples, we consider that any numbers a and b satisfy $a \ll b$ if $a/b \leq 1/10$. In case 3 our assumption $1/\sqrt{u_2} \ll n \ll 1/u_1$ (1) about intermediate population size holds, and we can see that the simulated mean is very close to the predicted mean. In cases 1 and 5, we replace the upper and lower inequalities in (1) by equality, respectively, so in each case one of the two assumptions is not valid. Cases 2 and 4 are

intermediate, meaning that the upper and lower inequalities in (1), respectively, don't quite hold since the ratios are 1/4. Yet, cases 2 and 4 show good agreement with the predicted mean.

The last two cases are specific examples related to regulatory sequence evolution in Drosophila and humans, which we will consider in more detail in the next section. The Drosophila effective population size is too large to use the true value in the simulations, but this is feasible for humans. In addition, we multiply u_2 by 1/3 for these special cases so that the ratio $u_1/u_2 = 30$. This comes from our assumption that a mutation at any position in the binding site will damage it, but to create a new binding site we require one position to mutate to the correct letter. Given a 10 letter target binding site, then u_1 is 3×10 times bigger than u_2.

In Figure 2, we plot the observed waiting time/predicted mean for case 3 and see a good fit to the exponential distribution, which agrees with our theoretical prediction. Figure 3 corresponds to case 1 and shows that the tail of the distribution looks exponential, but the simulated mean time is roughly 1.5 times larger than the predicted mean. This is caused by the fact that since $u_1 = 1/(2N)$, the time $\tau_B - \sigma_B$ we have to wait for the B mutant to be produced is of the same order of magnitude as σ_B. Hence, the total waiting time τ_B is significantly larger than σ_B.

To explain the observed shape of the distribution, recall from the sketch of the proof of Theorem 1 that σ_B has exactly an exponential distribution. Adding the independent random variable $\tau_B - \sigma_B$, which we will assume has density $g(s)$, yields the following distribution for τ_B:

$$\int_0^t e^{-(t-s)} g(s) ds = e^{-t} \int_0^t e^s g(s) ds \approx t e^{-t} g(0) \quad \text{when } t \text{ is small}$$

since the integrand is close to $g(0)$ for all $s \in [0, t]$, and consequently, the exponential fit is

not good for small t. WODARZ and KOMAROVA (2005) have done an exact calculation of the waiting time in the branching process approximation of the Moran model, which applies to case 1. As Figure 3 shows, the computation matches the Moran model simulation exceptionally well.

Figure 4 corresponds to case 5 and yields a simulated mean of about 78% of its predicted value. The curve looks exponential, but it has the incorrect mean. In this case, Theorem 3 shows that fixation of an A mutation and stochastic tunneling are both possible scenarios in which a B mutation can arise, producing a shorter waiting time. More specifically, the assumptions of Theorem 3 hold with $\gamma = 1$ which gives $\alpha = 1.433$, and Figure 3 shows that the exponential distribution with this rate fits the simulated data reasonably well.

EXPERIMENTAL RESULTS

In the following two examples, we consider a 10 nucleotide binding site and suppose that transcription factor binding requires an exact match to its target. We assume that any mutation within the binding site will damage it (A mutation) and that there exists at least one 10 nucleotide sequence within the regulatory region that can be promoted to a new binding site by one mutation (B mutation). Our previous results show, as MACARTHUR and BROOKFIELD (2004) had earlier observed, that the existence of these so-called "presites" is necessary for the evolution of new binding sites on a reasonable time scale (DURRETT and SCHMIDT 2007).

Drosophila: We will assume a per nucleotide mutation rate of 10^{-8} per generation, a simplification of the values that can be found in the classic paper of DRAKE *et al.* (1998) and the recent direct measurements of HAAG-LIAUTARD *et al.* (2007). If transcription factor binding involves an exact match to a 10 nucleotide target then inactivating mutations have probability $u_1 = 10^{-7}$ and those that create a new binding site from a 10 letter word

that doesn't match the target in one position have probability $u_2 = 1/3 \times 10^{-8}$. If the target word is 6–9 nucleotides long or inexact matches are possible, then these numbers may change by a factor of two or three. Such details are not very important here, since our aim is to identify the order of magnitude of the waiting time.

We will set the effective population size $N = 2.5 \times 10^6$ which agrees with the value given on page 1612 of THORNTON and ANDOLFATTO (2006). In order to apply Theorem 1 we need

$$1/\sqrt{u_2} = 1.73 \times 10^4 \quad \ll \quad 2N = 5 \times 10^6 \quad \ll \quad 1/u_1 = 10^7$$

The ratio of the left number to the middle number is $\approx 1/300$, but the ratio of the middle number to the one on the right is $1/2$ which says that $2N \ll 1/u_1$ is not a valid assumption. Ignoring this for a moment, Theorem 1 predicts a mean waiting time of

$$\frac{1}{2Nu_1\sqrt{u_2}} = \frac{1.73}{5} \times 10^{-6+7+4} \approx 34,600 \text{ generations}$$

which translates into 3,460 years if we assume 10 generations per year.

Since the assumption $2N \ll 1/u_1$ is not valid, we use our simulation result for case 6, which has a small population size with parameter values of $2Nu_1 = 0.5$ and $2N\sqrt{u_2} = 10/\sqrt{3} = 5.77$ similar to the Drosophila example, to see what sort of error we expect (see Tables 1 and 2). We see that in the simulation, the observed mean is approximately 25% higher than the theoretical mean, so adding 25% to the prediction gives a mean waiting time of 4,325 years.

A second and more important correction to our prediction is that Theorem 1 assumes that the A mutation is neutral and the B mutation is strongly advantageous. If we make the conservative assumption that the B mutation is neutral then the fixation probability $\beta = 1/2N = 2 \times 10^{-7}$, and by Theorem 1 the waiting time increases by a factor of $1/\sqrt{\beta} \approx 2200$

to about 9 million years. If the B mutation is mildly advantageous, i.e., $s - 1 = 10^{-4}$, then $\beta \approx 10^{-4}$ and the waiting time increases only by a factor of 100 to 400,000 years.

If we assume that A mutants have fitness $r < 1$ where $1 - r \ll \sqrt{u_2} = 5.78 \times 10^{-5}$ then Theorem 4 implies that the waiting time is not changed but if $(1 - r)/\sqrt{u_2} = \rho$ then the waiting time is increased by a factor $2/(\sqrt{\rho^2 + 4} - \rho) \approx \rho$ if ρ is large. If we use the value of $1 - r = 10^{-4}$ the increase is roughly a factor of 2. From this we see that if both mutations are almost neutral (i.e. relative fitnesses $r \approx 1 - 10^{-4}$ and $s \approx 1 + 10^{-4}$), then the switch between two transcription factor binding sites can be done in less than a million years. This is consistent with the results for the *even-skipped* stripe 2 enhancer mentioned earlier.

Humans: We will now show that two coordinated changes that turn off one regulatory sequence and turn on another without either mutant becoming fixed are unlikely to occur in the human population. We will assume a mutation rate of 10^{-8}, again see DRAKE *et al.* (1998), and an effective population size of $N = 10^4$ because this makes the nucleotide diversity $4N_e\mu$ close to the observed value of 0.1%. If we again assume that transcription factor binding involves an exact match to a 10 nucleotide target then inactivating mutations have probability $u_1 = 10^{-7}$, and those that create a new binding site from a 10 letter word that doesn't match the target in one position have probability $u_2 = 3.3 \times 10^{-9}$. For the assumptions of Theorem 1 to be valid we need

$$1/\sqrt{u_2} = 1.73 \times 10^4 \quad \ll \quad 2N = 2 \times 10^4 \quad \ll \quad 1/u_1 = 10^7$$

The ratio of the middle number to the one on the right is 1/500, but the ratio of the left number to the middle one is ≈ 1.

Ignoring for the moment that one of the assumptions is not satisfied, Theorem 1 predicts

a mean waiting time of

$$\frac{1}{2Nu_1\sqrt{u_2}} = \frac{1.73}{2} \times 10^7 = 8.66 \times 10^6 \text{ generations}$$

Multiplying by 25 years per generation gives 216 million years.

As shown in Tables 1 and 2, we have simulation results for humans using the exact parameters above. In 10,000 replications, the simulation mean is 6.46 million generations which is only about 75% of the predicted value. Multiplying by 0.75 reduces the mean waiting time to 162 million years, still a very long time. Our previous work has shown that, in humans, a new transcription factor binding site can be created by a single mutation in an average of 60,000 years, but, as our new results show, a coordinated pair of mutations that first inactivates a binding site and then creates a new one is very unlikely to occur on a reasonable time scale.

To be precise, the last argument shows that it takes a long time to wait for two prespecified mutations with the indicated probabilities. The probability of a 7 out of 8 match to a specified 8 letter word is $8(3/4)(1/4)^7 \approx 3.7 \times 10^{-4}$, so in a 1 kb stretch of DNA there is likely to be only one such match. However, LYNCH (2007, see page 805) notes that transcription factor binding sites can be found within a larger regulatory region ($10^4 - 10^6$ bp) in humans. If one can search for the new target sequence in $10^4 - 10^6$ bp, then there are many more chances. Indeed since $(1/4)^8 \approx 1.6 \times 10^{-5}$, then in 10^6 bp we expect to find 16 copies of the 8 letter word.

The edge of evolution? Our final example of waiting for two mutations concerns the emergence of chloroquine resistance in *P. falciparum*. Genetic studies have shown, see WOOTON *et al.* (2002), that this is due to changes in a protein PfCRT, and that in the mutant strains two amino acid changes are almost always present - one switch at position 76 and another at position 220. This example plays a key role in the chapter titled "The

Mathematical Limits of Darwinism" in Michael Behe's book, *The Edge of Evolution*.

Arguing that (i) there are a trillion parasitic cells in an infected person, (ii) a billion infected persons on the planet, and (ii) chloroquine resistance has only arisen ten times in the last fifty years, he concludes that the odds of one parasite developing resistance to chloroquine, an event he calls a *chloroquine complexity cluster* or CCC, is roughly 1 in 10^{20}. Ignoring the fact that humans and *P. falciparum* have different mutation rates he then concludes that "On the average, for humans to achieve a mutation like this by chance, we would have to wait a hundred million times ten million years," which is five million times larger than the calculation we have just given.

Indeed his error is much worse. To further sensationalize his conclusion, he argues that "There are 5000 species of modern mammals. If each species had an average of a million members, and if a new generation appeared each year, and if this went on for two hundred million years, the likelihood of a single CCC appearing in the whole bunch over that entire time would only be about 1 in 100." Taking $2N = 10^6$ and $\mu_1 = \mu_2 = 10^{-9}$, Theorem 1 predicts a waiting time of 31.6 million generations for one prespecified pair of mutations in one species, with $\sqrt{u_2}$ having reduced the answer by a factor of 31,600.

We are certainly not the first to have criticized Behe's work. LYNCH (2005) has written a rebuttal to BEHE and SNOKE (2004), which is widely cited by proponents of intelligent design (see the Wikipedia article on Michael Behe). BEHE and SNOKE (2004) consider evolutionary steps that require changes in two amino acids and argue that to become fixed in 10^8 generations would require a population size of 10^9. One obvious problem with their analysis is that they do their calculations for $N = 1$ individual ignoring the population genetics effects that produce the factor of $\sqrt{u_2}$. LYNCH (2005) also raises other objections.

CONCLUSIONS

For population sizes and mutation rates appropriate for Drosophila, a pair of mutations can switch off one transcription factor binding site and activate another on a time scale of several million years, even when we make the conservative assumption that the second mutation is neutral. This theoretical result is consistent with the observation of rapid turnover of transcription factor binding sites in Drosophila and gives some insight into how these changes might have happened. Our results show that when two mutations with rates u_1 and u_2 have occurred and

$$1/\sqrt{u_2} \ll 2N \ll 1/u_1$$

then the first one will not have gone to fixation before the second mutation occurs, and indeed A mutants will never be more than a small fraction of the overall population. In this scenario, the A mutants with fitness r are significantly deleterious if $(1-r)/\sqrt{u_2}$ is large, a much less stringent condition than the usual condition that $2N(1-r)$ is large. Also, the success probability of the B mutant is dictated by its fitness relative to the wild type rather than relative to the A mutant. This follows because the fraction of A mutants in the population is small when the B mutant arises, and hence most individuals are wild type at that time.

The very simple assumptions we have made about the nature of transcription factor binding and mutation processes are not crucial to our conclusions. Our results can be applied to more accurate models of binding site structure and mutation processes whenever one can estimate the probabilities u_1 and u_2. However, the assumption of a homogeneously mixing population of constant size is very important for our analysis. One obvious problem is that Drosophila populations undergo large seasonal fluctuations, providing more opportunities for mutation when the population size is large, and a greater probability of fixation of an A mutation during the recurring bottlenecks. Thus, it is not clear that one can reduce to a

constant size population, or that the effective population size computed from the nucleotide diversity is the correct number to use for the constant population size. A second problem is that in a subdivided population, A mutants may become fixed in one subpopulation giving more opportunities for the production of B mutants, or perhaps leading to a speciation event. It is difficult to analyze these situations mathematically, but it seems that each of them would increase the rate at which changes occur. In any case one would need to find a mechanism that changes the answer by a significant factor in order to alter our qualitative conclusions.

ACKNOWLEDGEMENTS

We would like to thank Eric Siggia for introducing us to these problems and for many helpful discussions and Jason Schweinsberg for his collaboration on related theoretical results. We are also grateful to Nadia Singh, Jeff Jensen, Yoav Gilad and Sean Carroll who commented on previous drafts, and two referees (one anonymous and one named Michael Lynch) who helped to improve the presentation. Both authors were partially supported by a National Science Foundation/National Institutes of General Medical Sciences grant (DMS 0201037). R.D. is also partially supported by a grant from the probability program (DMS 0202935) at the National Science Foundation. D.S. was partially supported by a National Science Foundation graduate fellowship at Cornell, and after graduation was postdoc at the Institute for Mathematics and its Applications in Minnesota, 2007–2008.

LITERATURE CITED

Behe, M., 2007 *The Edge of Evolution. The search for the limits of Darwinism.* Free Press, New York.

Behe, M., and D. W. Snoke, 2004 Simulating evoluton by gene duplication of protein features that require multiple amino acid residues. Protein Science **13:** 2651–2664.

Carter, A. J. R., and G. P. Wagner, 2002 Evolution of functionally conserved enhancers can be accelerated in large populations: a population-genetic model. Proc. R. Soc. London **269:** 953–960.

Drake, J. W., B. Charlesworth, D. Charlesworth, and J. F. Crow, 1998 Rates of spontaneous mutation. Genetics **148:** 1667–1686.

Durrett, R., and D. Schmidt, 2007 Waiting for regulatory sequences to appear. Annals of Applied Probability **17:** 1-32.

Durrett, R., D. Schmidt, and J. Schweinsberg, 2007 A waiting time problem arising from the study of multi-stage carcinogenesis. *Annals of Applied Probability*, to appear. Manuscript available at `www.math.cornell.edu/~durrett`

Ewens, W. J., 2004 *Mathematical Population Genetics.* Second Edition. Springer-Verlag, New York.

Haag-Liautard, C., M. Dorris, X. Maside, S. Macskill, D. L. Halligan, B. Charlesworth, and P. D. Keightley, 2007 Direct estimation of per nucleotide and genomic deleterious mutation rates in Drosophila. Nature **445:** 82–85.

Iwasa, Y., F. Michor, and M. A. Nowak, 2004 Stochastic tunnels in evolutionary dynamics. Genetics **166:** 1571–1579.

Iwasa, Y., F. Michor, N. L. Komarova, and M. A. Nowak, 2005 Population genetics of tumor suppressor genes. J. Theor. Biol. **233:** 15–23.

Kimura, M., 1985 The role of compensatory mutations in molecular evolution. *J. Genet.* **64:** 7–19.

Kimura, M., and T. Ohta, 1969 The average number of generations until the fixation of a

mutant gene in a finite population. Genetics **61:** 763–771.

Komarova, N. L., A. Sengupta, and M. A. Nowak, 2003 Mutation-selection networks of cancer initiation: tumor suppressor genes and chromosomal instability. J. Theor. Biol. **223:** 433–450.

Ludwig, M. Z., N. H. Patel, and M. Kreitman, 1998 Functional analysis of eve stripe 2 enhancer evolution in Drosophila: rules governing conservation and change. Development **125:** 949–958.

Ludwig, M. Z., C. E. Bergman, N. H. Patel, and M. Kreitman, 2000 Evidence for stabilizing selection in a eukaryotic enhancer element. Nature **403:** 564–567.

Ludwig, M. Z., A. Palsson, E. Alekseeva, C. E. Bergman, J. Nathan, and M. Kreitman, 2005 Functional evolution of a *cis*-regulatory module. PLoS Biology **3:** 588–598.

Lynch, M., 2005 Simple evolutionary pathways to complex proteins. Protein Science **14:** 2217–2225.

Lynch, M., 2007 The evolution of genetic networks by non-adaptive processes. Nature Reviews Genetics **8:** 803–813.

MacArthur, S., and J. F. Brookfield, 2004 Expected rates and modes of evolution of enhancer sequences. Mol. Biol. Evol. **21:** 1064–1073.

Nowak, M. A., 2006 *Evolutionary Dynamics: Exploring the Equations of Life.* Belknap Press, Cambridge, MA.

Stone, J. R., and G. A. Wray, 2001 Rapid evolution of *cis*-regulatory sequences via local point mutations. Mol. Biol. Evol. **18:** 1764–1770.

Thornton, K., and P. Andolfatto, 2006 Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherland population of *Drosophila melanogaster*. Genetics **172:** 1607–1619.

Wooton, J. C., Feng, A., Ferdig, M. T., Cooper, R. A., Mu, J., et al., 2002 Genetic diversity and chloroquine selective sweeps in *Plasmodium falciparum. Nature.* **418:** 320–323.

Wodarz, D., and N. L. Komarova, 2005 *Computational Biology of Cancer: Lecture notes and mathematical modeling.* World Scientific Publishing, Hackensack, NJ.

Parameters

	u_1	$\sqrt{u_2}$
Case 1	$1/n$	$10/n$
Case 2	$1/4n$	$10/n$
Case 3	$1/10n$	$10/n$
Case 4	$1/10n$	$4/n$
Case 5	$1/10n$	$1/n$
Drosophila	$1/2n$	$10/\sqrt{3}n$
Humans	$2/1000n$	$2/\sqrt{3}n$

Table 1: Parameter values for our simulations in terms of $n = 2N$.

Comparison of mean waiting times

	Population Size	Predicted Mean	Simulated Mean	Sim/Pred
Case 1	1,000	100	156.5	1.565
	10,000	1,000	1,559	1.559
Case 2	1,000	400	460.0	1.150
	10,000	4,000	4,552	1.138
Case 3	1,000	1,000	1,047	1.048
	10,000	10,000	10,414	1.041
Case 4	1,000	2,500	2,492	0.997
	10,000	25,000	2,5271	1.011
Case 5	1,000	10,000	7,811	0.781
	10,000	100,000	77,692	0.777
Drosophila	1,000	346.4	441.1	1.273
	10,000	3,464	4,374	1.263
Humans	20,000	8,660,258	6,464,920	0.747

Table 2: Mean waiting times for $n = 1,000$ and $n = 10,000$ with mutation rates u_1 and u_2 as defined by each case. The simulated mean time is compared with the predicted mean from Theorem 1 and the ratio is given in the last column. All simulations are done with 10,000 replications.

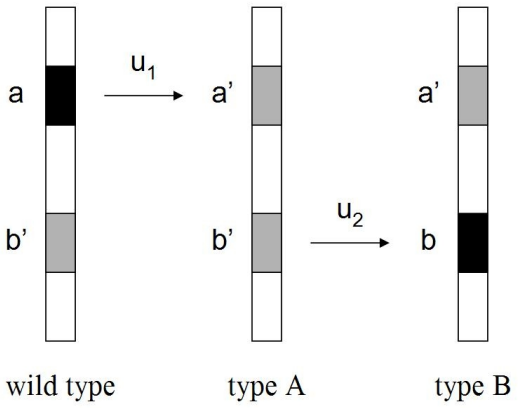

Figure 1: An example of our general two stage mutation process used in this paper is as follows. The regulatory region contains two possible binding sites, a and b, where a prime denotes an inactivated site. Wild type individuals can undergo a type A point mutation (a b' $\rightarrow$ a' b' at rate u_1) which inactivates site a, and type A individuals can undergo a type B point mutation (a' b' $\rightarrow$ a' b at rate u_2) which creates a new active site b. The relative fitnesses of wild type, A mutant, and B mutant are 1, r, and s, respectively. Note that in this case, wild type individuals cannot produce individuals with a second active binding site. For a different example of this general process, see page 955 of CARTER and WAGNER (2002).

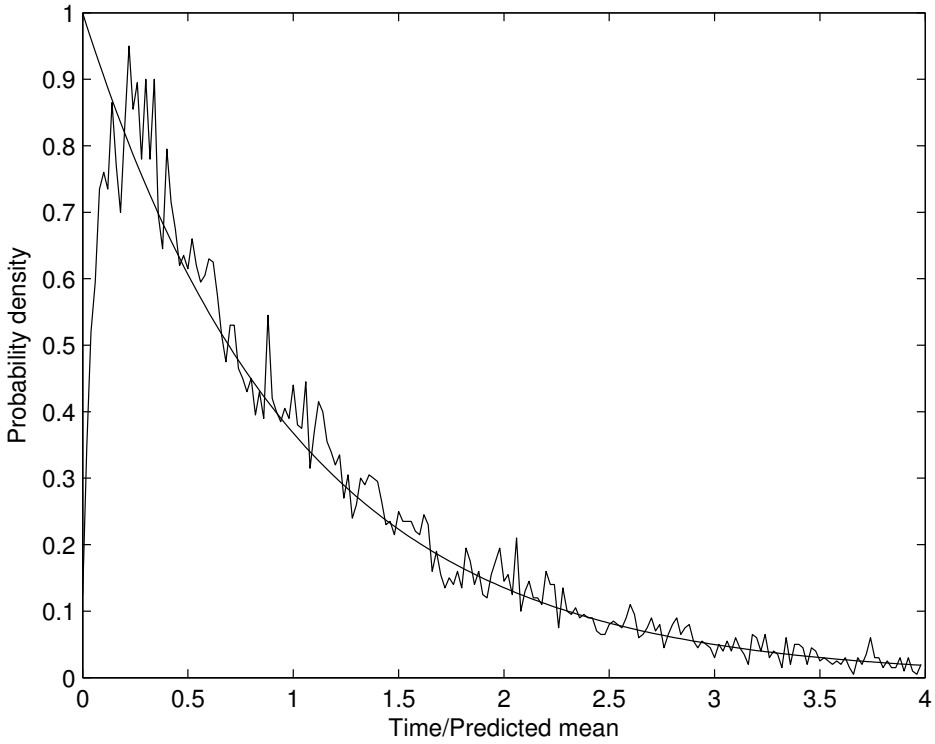

Figure 2: Waiting time for case 3 with $n = 1,000$ and $u_1 = u_2 = 0.0001$. The assumptions for intermediate population size as compared to mutation rates (1) hold and, as predicted by Theorem 1, the waiting time is a good fit to the exponential distribution.

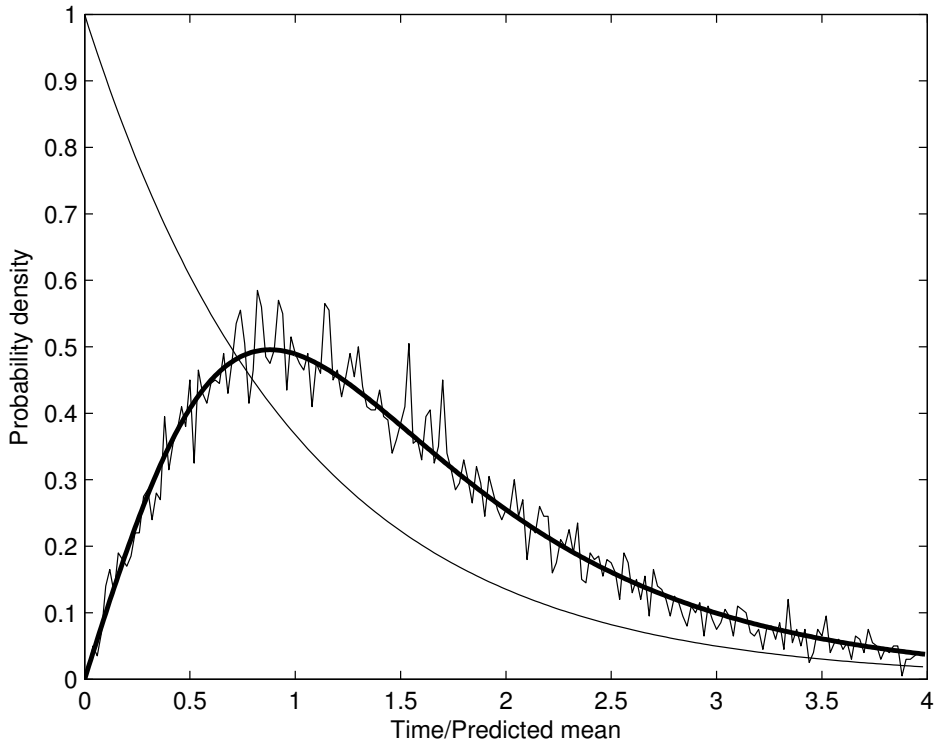

Figure 3: Waiting time for case 1 with $n = 1,000$, $u_1 = 0.001$, and $u_2 = 0.0001$. The tail of the waiting time distribution appears to be exponential, but the simulated mean is about 1.5 times larger than predicted by Theorem 1. The thick curve corresponds to the waiting time calculation done by Wodarz and Komarova (2005). This computation matches our Moran model simulation exceptionally well.

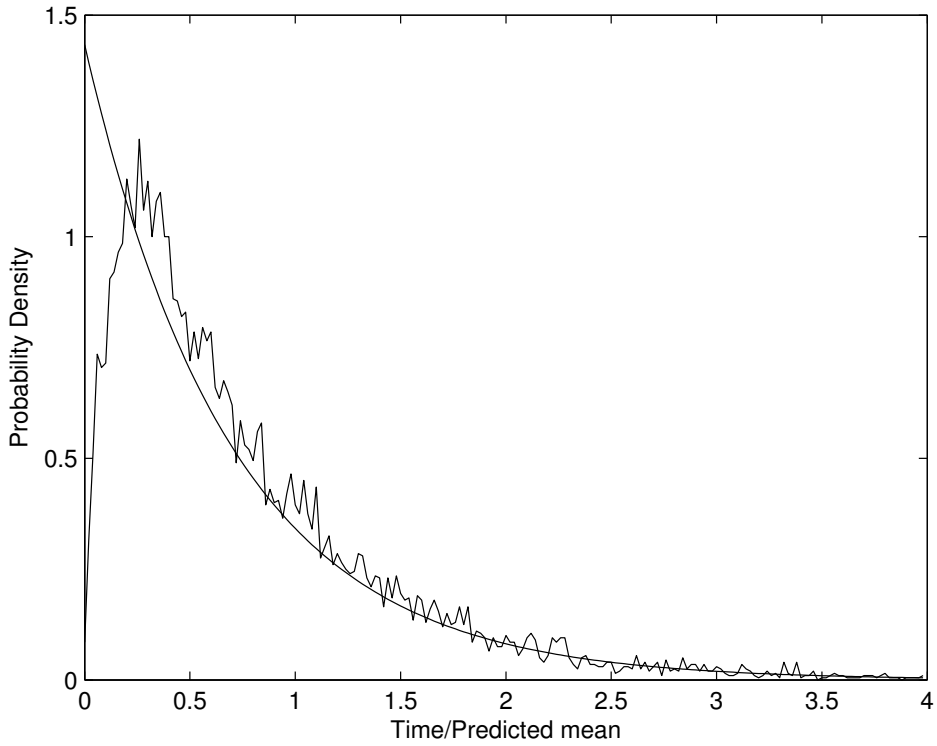

Figure 4: Waiting time for case 5 with $n = 1,000$, $u_1 = 0.0001$, and $u_2 = 0.000001$. The simulated mean is only about 78% of the mean predicted by Theorem 1, however, the waiting time distribution still is approximately exponential. Theorem 3 holds with $\gamma = 1$ so $\alpha = 1.433$, and the exponential with this rate gives a reasonable fit to the simulated data.