Mutate now, die later evolutionary dynamics with delayed selection

Andreea Munteanu, Peter F. Stadler

www.elsevier.com/locate/yjtbi

Mutate now, die later evolutionary dynamics with delayed selection

Andreea Munteanu, Peter F. Stadler

PII: S0022-5193(09)00299-9
DOI: doi:10.1016/j.jtbi.2009.06.022
Reference: YJTBI 5612

To appear in: *Journal of Theoretical Biology*

Received date: 15 February 2009
Revised date: 15 June 2009
Accepted date: 24 June 2009

Cite this article as: Andreea Munteanu and Peter F. Stadler, Mutate now, die later evolutionary dynamics with delayed selection, *Journal of Theoretical Biology*, doi:10.1016/j.jtbi.2009.06.022

Mutate Now, Die Later.
Evolutionary Dynamics with Delayed Selection

Andreea Munteanu[*,a], Peter F. Stadler[b,1,d,e,f]

[a] *ICREA-GRIB Complex Systems Lab, UPF, Parc de Recerca Biomedica Barcelona Dr Aiguader 88, E-08003 Barcelona, Spain*
[b] *Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany*
[c] *Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, D-04103 Leipzig, Germany*
[d] *Fraunhofer Institut für Zelltherapie und Immunologie – IZI Perlickstraße 1, D-04103 Leipzig, Germany*
[e] *Department of Theoretical Chemistry University of Vienna, Währingerstraße 17, A-1090 Wien, Austria*
[f] *Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA*

Abstract

We analyze here the evolutionary consequences of selection with delay in a population genetics context. In the classical works on evolutionary dynamics, an individual produces off-springs in direct proportion to its fitness, a process in which mutations may occur. In the present scenario of delayed selection, individuals that acquire deleterious mutations can still reproduce unharmed for several generations. During this time delay, the damage passed on to off-springs can potentially be repaired by subsequent compensatory mutations. In the absence of such a repair, the individual becomes sterile. Here we study the population-genetic effects of such a time delay by means of both numerical simulations and theoretical modeling. The results show that delayed selection lowers the extinction threshold, endangering the survival of the population. Surprisingly, however, no traces of this delay effect are encountered in the sequence diversity of the population. These conclusions suggest that delayed selection is hard to detect in genetic data and thus could be a wide-spread but rarely detected phenomenon.

Key words: Evolution, neutral networks, fitness, extinction, telomere

1. Introduction

Darwinian evolution is the interplay of the production of variation and subsequent selection. Due to the complexity of biological organism, selection tends to act at all times, punishing or rewarding small differences among individuals. This is not necessarily the case at the level of (small) genetic subsystems, however. The intuitive rationale for this claim is that an "emergency subsystem", for instance, may not need to be activated for several generations. While unused and inactive, it tends to escape the forces of selection and conceivably, acquire damages. Once conditions change and it is needed again, however, there are severe (fitness) penalties if its functionality has not been maintained or repaired. We expect such "delayed selection" to leave detectable traces in the genome. Hence we study here the dynamical implications of delayed selection in some detail.

It may come as a surprise that the best studied example is a generic component of the eukaryotic replication machinery, namely the reconstruction of telomere ends. Mice deficient for the mouse telomerase RNA (mTR-/-) are fertile and show initially little if any pathologies. However, they can breed only for about six generations due to decreased male and female fertility and to an increased embryonic lethality in later generations. Even late generation (mTR-/-) mice are viable to adulthood, only showing a decrease in viability in old age (Lee et al., 1998; Herrera et al., 1999). These effects appear to be linked to the shortening of the telomeres (Verdun and Karlseder, 2007). Similar effects can be observed in cell culture, again establishing a relationship between viability and telomere length: Terc-deficient embryonic stem cells show gradual reduction of growth rate after about 300 divisions, and proliferation virtually stops after 450 generations (Niida et al., 1998). At the same

[*]Corresponding author at: Parc de Recerca Biomedica Barcelona Dr Aiguader 88, E-08003 Barcelona, Spain
Email addresses: andreea.munteanu@upf.edu (Andreea Munteanu), studla@bioinf.uni-leipzig.de (Peter F. Stadler)

time, telomerase RNA exhibits extremely high rates of evolution (Xie et al., 2008). The speculation that delayed selection may be part of the explanation for the unexpected evolutionary plasticity of telomerase RNA motivated this work.

Delayed selection is also likely to occur in species for which environmental conditions vary periodically at timescales longer than generation time. A spectacular example is the monarch butterfly (*Danaus plexippus*) (Urquhart, 1960). The "migratory" generation migrates from Eastern North America to overwintering sites in Mexico. This long-lived generation is characterized by reproductive diapause persisting until next spring, when the butterflies reproduce and start the journey back north. Another two to three generations of reproductively competent, short-lived "summer" butterflies follow the progressive, northward emergence of milkweed. Significant differences in gene expression between summer and migratory butterflies (Zhu et al., 2008) suggest that some parts of the butterflies genetic system may be unused over a few generations. Whether this is indeed the case could be tested directly if characteristic genomic fingerprints of delayed selection can be detected.

A more subtle context in which delayed selection may play a role is that of synthetically lethal genes. A pair (or a larger set) of genes is called *synthetically lethal* if knocking out the entire set is lethal, while the knockout of all smaller subsets retains viability (Hartman IV et al., 2001; Kaelin Jr, 2005; Le Meur and Gentleman, 2008). Note that synthetically lethal gene pairs typically share their primary function but cannot be redundant in all their functional aspects. The reason is that exact redundancy is evolutionarily unstable: it is quickly resolved by the loss of one copy (Force et al., 1999). This type of genetic buffering may, however, delay the detrimental effects of functional loss in one partner until a rarely employed secondary function of the affected gene is required. Again, a recognizable signal in the genomic DNA would be of utmost interest.

The paper is organized as follows: in section §2 we introduce the methodology and the results of the stochastic simulations for a population of RNA molecules. In section §3, we confront these results with a mean-field model that captures the evolution of the population in a delayed-selection scenario. We quantify the amount of diffusion in the sequence space for various time delays in search of a signature on the evolutionary rates of such altered selection. Finally, we discuss the findings, with special emphasis on the lack of such an inequivocable signature, in the context of genomic studies in section §4.

2. RNA-Based Simulations

2.1. A Simple Model of Telomere Damage

The simulation framework used in this contribution is motivated by the telomerase RNA (TR) system briefly discussed in the introduction. For simplicity we distinguished only between fitness-neutral and lethal mutations. Each individual is characterized by its TR gene and the length of its telomere. Off-springs with intact TR have full-length telomeres, while telomeres shrink by a constant amount with each replication step in which the telomerase is inactive. Individuals whose telomeres have shrunk to zero are sterile, i.e., their fitness is set to 0.

In order to include a genetic component with a realistic genotype-phenotype map, we use RNA secondary structures to represent phenotypes. In this approach, each sequence s is folded into its minimum energy secondary structure $\varphi(s)$ and then fitness is evaluated by comparing $\varphi(s)$ with a target structure φ^*, (see e.g. Fontana et al. 1989; Schuster et al. 1994; Huynen et al. 1996a). Here, we stipulate that only the target secondary structure is functional. The fitness f of an individual with genotype s and telomere length k is given thus by

$$f(s,k) = \begin{cases} 1 & \text{if } \varphi(s) = \varphi^* \text{ or } k > 0 \\ 0 & \text{if } \varphi(s) \neq \varphi^* \text{ and } k = 0 \end{cases} \qquad (1)$$

Since the computational effort for RNA folding computations is cubic in sequence length, vertebrate TR gene with 300-500nt are too long to practical for our simulations. Instead of a real TR structures, we defined an arbitrarily chosen target structure of length 100 to represent the viable phenotype. RNA secondary structure predictions are performed using the `Vienna RNA Package` (Hofacker et al., 1994).

We simulate a population of N individuals in a flow reactor under stochastically controlled constant organization as described in e.g. Fontana et al. (1989). Individuals replicate proportional to their fitness. During replication, each letter is mutated with a probability μ. Then the structure $\varphi(s')$ of the offspring s' is computed. If $\varphi(s') = \varphi^*$, we set $k' = K$, otherwise $k' = k' - 1$, where K is the number of generations for which a defective TR is tolerated. In other words, if after K replications, such an incorrect fold has not encountered the neutral network, its fitness becomes 0 and thus loses the capacity of replication.

In the following, we shall discuss the results of the simulations. Based on these data, we introduce a theoretical model associated to the simulation framework,

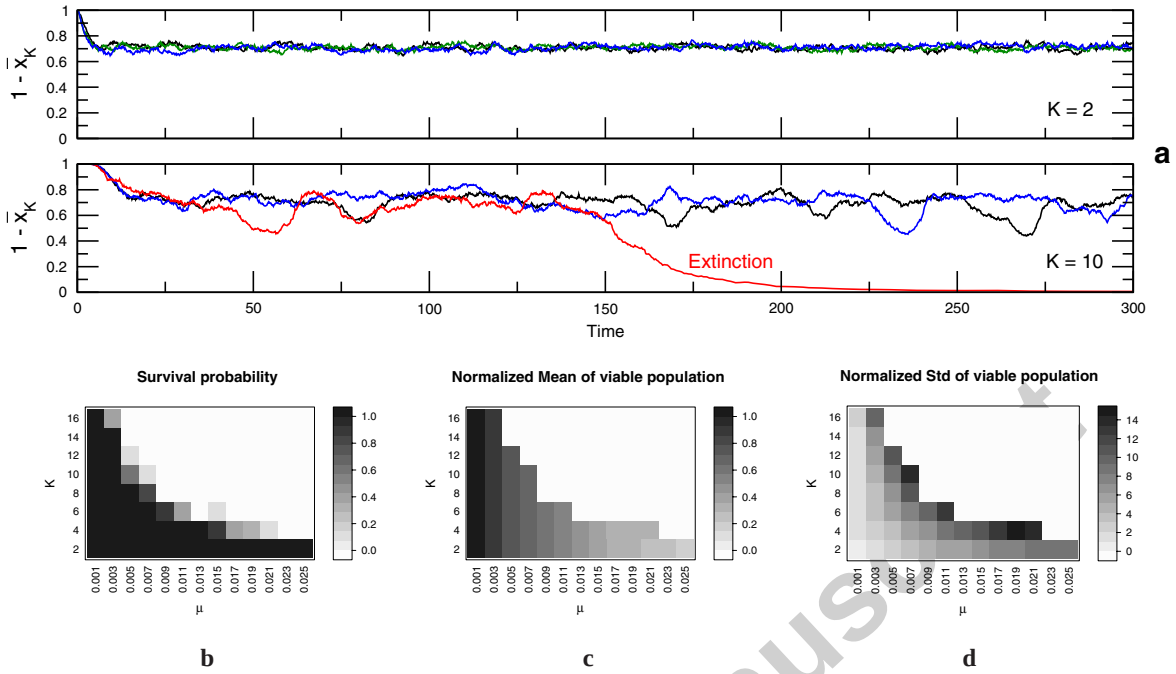

Figure 1: **Top** Time series of the mean fitness or fraction of reproducing individuals as a function of time in three stochastic runs with $N = 1000$, $\mu = 0.005$ and two different values of K. **Below:** Survival probability, time average and standard deviation over the time-series of the mean fitness as a function of K and μ. Averages are taken over 10 independent simulations running for 100,000 mutations for each combination of parameters. Average and standard deviation are normalized to their values at $K = 2$ and $\mu = 0.001$ to facilitate comparison.

a model that provides a reasonably-good fit of the simulations results and also a tool to better understand the implications of the delayed-selection effect.

2.2. Extinction Threshold

As a first observation, we notice that one of the consequences of the delayed selection is a reduced critical value of the mutation rate at which the population goes extinct. An erroneously replicating haploid population shows the so-called *error threshold* phenomenon, by which the population loses coherence and quickly approaches a uniform distribution in sequence space as soon as the mutation rate exceeds a critical value. Originally described on single-peak landscapes (Eigen, 1971; Eigen et al., 1989), an analogous phenomenon can be observed at the phenotypic level (Forst et al., 1995; Huynen et al., 1996a; Wilke, 2001). With instantaneous fitness effects, the critical value of μ can be estimated from a μ-dependence of the equilibrium concentration of the "poor" phenotypes.

Before we comment in detail the results obtained for the current framework, we wish to discuss the distinction between error threshold and extinction threshold, a distinction often disregarded in the literature. Extinction can be the consequence of a process such as lethal mutagenesis (Bull et al., 2007), with the latter being a demographic process occurring, for example, in the context of within-host population of viruses that become extinct with an elevated mutation rate. In this case, the population is overwhelmed by deleterious mutations and cannot sustain itself. Eigen's error catastrophe or error threshold, although inspired by the idea of lethal mutagenesis, is a distinct process. The error threshold is defined as the mutation rate beyond which the mean fitness of the population does not decrease exponentially with the mutation rate but remains constant, as all genotypes are insensitive to mutations (the information is lost from the population). Contrary to the intuition, Eigen two-class fitness landscape (population of only high- and low-fitness genotypes) actually retards population extinction (Bull et al., 2007). In the light of these comments, we witness in the current framework the process of stochastic extinction rather than an error threshold, and thus we refer to the mutation threshold as extinction threshold.

The upper panels of Fig. 1 represent the mean fit-

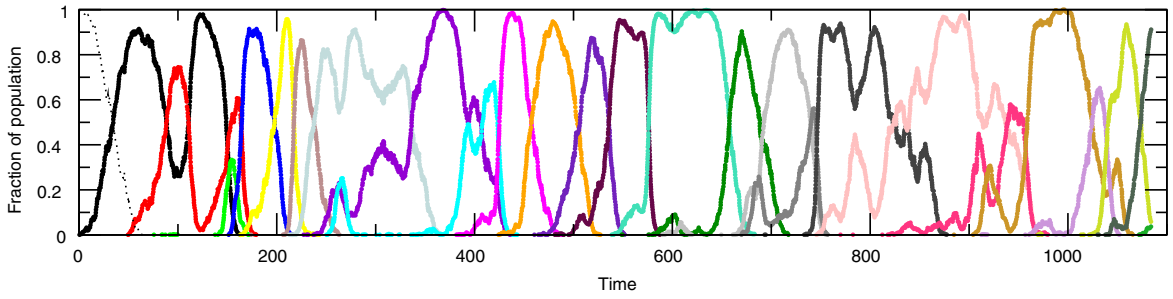

Figure 2: The time-series illustrating the dynamics of repairs in the population ($N = 1000, \mu = 0.005, K = 10$). The different colors represent waves of increasing number of repairs (up to 25 repairs) that sweep through the population as time increases. The dashed black line contains the initial population with no repairs.

ness (i.e. the fraction of reproducing or fit individuals) for several examples of simulations, showing that the stochastic extinction of the population at finite times is largely driven by an increase of the fluctuations. That is, for a fixed mutation rate μ, the average fraction of reproducing individuals is the same for various values of K, but the standard deviation increasing with K. For large K, due to these large excursions, the reproducing population may reach a threshold value at which extinction occurs. The main effect of delayed selection is thus a strong increase in fluctuations, that causes stochastic extinction in finite populations at mutation rates significantly lower than the non-delayed selection ($K = 1$). This can be seen in the lower panels of Fig. 1 where the extinction threshold or survival probability (panel *b*) is illustrated as resulting from the simulations. A rough estimation of the survival probability was considered to be the fraction of the simulations that have not gone extinct after a number $M = 100000$ of mutations. From these simulations, we have also estimated the dependency of the mean fitness (panel *c*) and its standard deviation (panel *d*). It can be seen that the mean fitness is not influenced by the telomere's length, while the fluctuations level (standard deviation) increases with the telomere's length. In §3 we shall pinpoint the causes of this premature extinction by means of a deterministic model.

2.3. Recoveries

The delayed selection has a direct effect on the extinction threshold in a negative way through the fluctuations described above, and in a positive one through the recoveries that might originate from damaged but still fit individuals.

In Fig. 2 we illustrate this effect by plotting the fraction of fit individuals characterized by a certain number of damage-and-repair cycles. It comes as a surprise that already after a very short time there are no lineages whose ancestry has had functional genes. The individuals without repairs (dotted black line) quickly disappear, as the ones with one damage-repair cycle (thick black line) appear, which in turn will be damaged and repaired again, transforming into the individuals with two cycles (red lines), and so on. We see that there is a characteristic time scale by which individual lineages acquire a damage and find their way back to the neutral network through subsequent mutations that repair the damage. That is, waves of repaired sequences swap the population, with newly repaired sequences displacing the old and less repaired sequences.

By comparing different simulations with identical (N, K, μ) values, we noticed that the stochastic effects dominate, i.e., there are dramatic fluctuations in the times between subsequent damage/repair events entirely due to stochasticity.

Through the simulations, we have measured the parameter R defined as the probability that a damaged telomerase recovers, i.e., the fraction of the replications occurring *off* the neutral network that give rise to an offspring *on* the neutral net through mutation/mutations. We expect the recovery fraction R not to depend on the length $K - i$ of the telomere, where $K - i$ also has the meaning of number of *replications* off the neutral net. From the stochastic simulations we see that this first approximation is acceptable, as it is illustrated in Fig. 3. For three experiments of equal K and different values of μ, we have recorded the number of recoveries and the type of sequences the recovery occurred from. The type of sequence refers to the length $K - i$ of the telomere and the number ℓ of *mutations* occurring off the neutral network. Naturally, at least two off-mutations are needed, one originating the fall off the neutral network

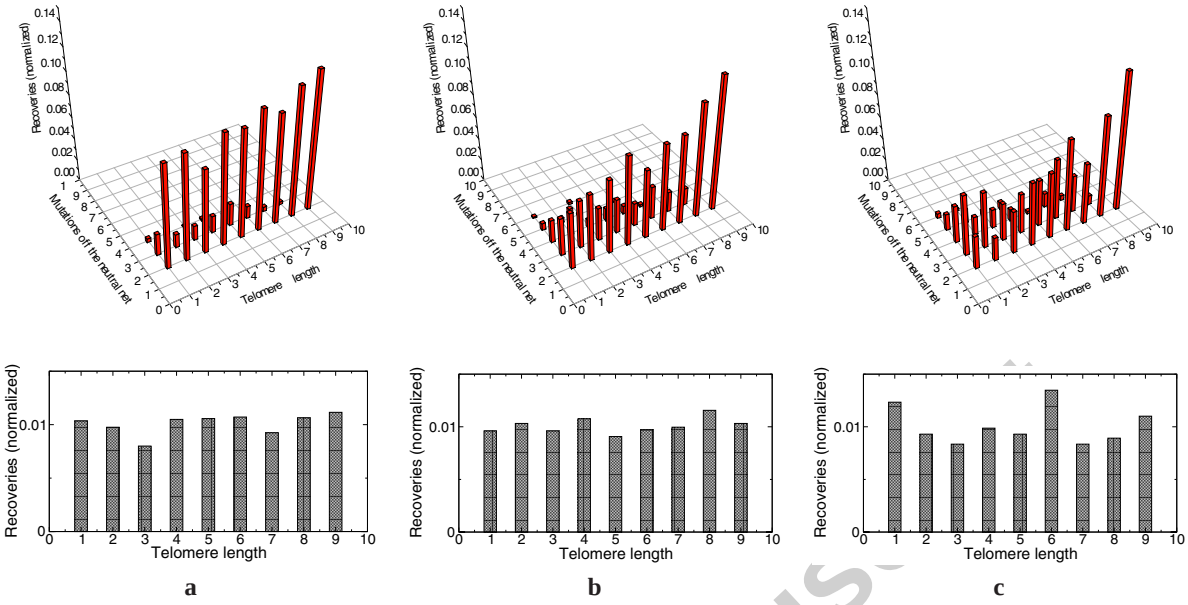

Figure 3: Distribution of recoveries $R_{i\ell}$ measured as a function of the telomere length $K - i$ and the number of off-network mutations ℓ. Simulation runs use $N = 1000$ and $K = 10$, and different mutation rates (a) $\mu = 0.001$, (b) $\mu = 0.003$, (c) $\mu = 0.005$. The lower panels show $R_i = \sum_\ell R_{i\ell}$, illustrating a reasonable independence on the telomere length, supporting the model from Fig. 5. Upper panels have been drawn using Dislin Scientific Plotting Software.

(the damage), and the second providing the recovery. In the lower panels we include the sum $R_i = \sum_\ell R_{i\ell}$, measuring thus the dependence of the recovery fraction on the telomere length alone, irrespective of the number of off-mutations needed. It can be seen that $R_i \equiv R$ is roughly independent on the telomere length (or number of replications occurring off the neutral networks), as the recovery mutation can occur with equal probability during the $K - 1$ replications prior to "death". As detailed in the next section, this observed independence allows us to construct a model of $K + 1$ variables, thus based only on the telomere length, without considering the number of off-mutations.

Returning to the notation $R_{i\ell}$, we find that, contrary to R_i, $R_\ell = \sum_i R_{i\ell}$ is not independent of the number of mutations off the neutral network ℓ, see Fig. 3. In fact, the recovery probability rapidly drops with ℓ. A more realistic model would take into account the $(K + 1) \times M$ variables, where $M \leq K$ is the maximal number of off-mutations. Implicitly, such a model would need to include more details on the structure of the neutral network and on the transition rates between neutral networks which could be borrowed from the presentation of Reidys et al. (2001). Since the purpose of this presentation is to provide a qualitative understanding of the consequences of delayed selection, we are content here with the much simpler, analytically tractable, case. It is worth noting in this context that the behavior of R is not an idiosyncratic property of RNA folding but rather a consequence of the generic properties of dense neutral networks. These can be modelled as dense and connected subgraphs of the sequence space (Reidys et al., 1997, 2001), a fact that accounts for the multiple paths of "repair" of structural damage, and hence relatively large values of R_1. Protein folding models (Babajide et al., 1997, 2001; Bastolla et al., 1999) show the same generic features.

From the stochastic simulations it can be seen that a rough estimate of this recovery fraction is on the order of 10% of the mutations occurring off the neutral network (Fig. 3). The repair or recovery of damaged genotypes by compensatory mutations thus has a dramatic effect on the long-time behavior of the population. To estimate the effect we shall introduce in §3 a model of the population dynamics which makes use of this observation of equal recovery fraction.

In addition, since R is defined as a conditional probability, we also expect that it will not depend strongly on the mutation rate μ for small values of μ. The parameter R will strongly depend on the size and structure

of the neutral network, and on its embedding the hypercube. This is the behavior followed also by the neutrality ν referring to the increased buffering, due to neutral networks, of the phenotype (the correct secondary structure) with respect to genetic mutations (nucleotide mutations) (Huynen et al., 1996b; Stadler et al., 2001). The strongest influence on the probability of recovery R is exerted by the distance of the mutant individual from the neutral network. As we shall see also in the modeling approach, an appropriate measure of the recovery rate or probability is defined through $\lambda = pR$, with p from eq. (4). This rate defines the probability that a replication leads to recovery. Considering a wide interval of μ-values, we expect $\lambda(\mu)$ to have an optimum for a certain value of $\mu_c \equiv \mu$. For $\mu \ll \mu_c$, the recovery probability is low as, once off the network, a new mutation is improbable to occur (p is small) in the next $K-1$ replications. For $\mu \gg \mu_c$, once off the neutral net, several nucleotides can mutate in a replication event, and thus destroy the repair. In addition, this regime of relatively large μ is limited by the extinction threshold, as we have seen that large fluctuations can lead the system into extinction.

In the context of these considerations, we have measured the recovery rate from the simulations. Based on the definitions introduced above, we monitored the temporal evolution of the number of repairs and that of replications with mutations occurring for sequences that are off the neutral network. The parameter R, the ratio of these two quantities, stabilizes after a transient period. We show these post-transient values from simulations for various μ and K in the upper panel of Fig. 4. Due to stochasticity, simulations of identical (N, K, μ) may lead to slightly different values of R. The dependence on μ is evident, as well as on K, with the former being more pronounced than the latter. in the lower panel, as commented above, an optimum value μ_c is apparent for which λ, the recovery probability, has a maximum. It is interesting though that lower K implies higher recovery probability. This can be explained by the structure of the neutral networks of RNA secondary structures. These are dense and fairly homogeneous only within the set of sequences that are compatible with the target structure (Schuster et al., 1994; Gruener et al., 1996; Sumedha et al., 2007; Jorg et al., 2008). Mutations that destroy compatibility (i.e., those that violate the base pairing rule), however, may lead away from the neutral network of the functional structure. Two or more incompatible substitutions therefore lead to regions in sequence space from where recovery in a single step is impossible.

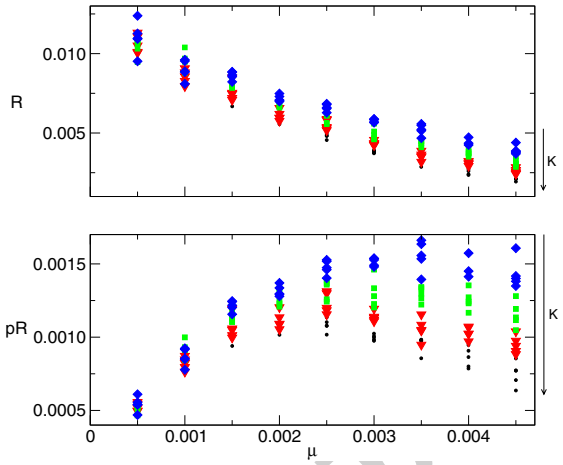

Figure 4: The recovery fraction R (a) and recovery rate $\lambda \equiv pR$ (b) as they result from 5 simulation runs of identical parameters. It illustrates their dependence on the mutation rate μ and K: $K = 10$, black circles; $K = 9$, red triangles; $K = 8$, green squares; $K = 7$, blue diamonds. The dependence on K becomes more pronounced at higher μ.

3. Deterministic Model

3.1. Replication Kinetics

Since we are interested in the basic effects of delayed selection, we neglect the influence of complex genetics and restrict ourselves to the simplest case of a population of haploid individuals. Naturally, this leads us to a variant of Eigen's Quasispecies Model (Eigen, 1971; Eigen et al., 1989). While certain issues, such as the influence of delay on the Extinction Threshold, could be studied in an even simpler setting, we explicitly include the redundancy of the genotype-phenotype mapping (Schuster et al., 1994). For simplicity, we only model the loss of fertility of individuals whose telomeres have disappeared. The population is structured into $K+1$ distinct groups of sequences characterized by a certain telomere length between 0 and K (Fig.5). We index these classes by the amount of telomere loss, so that x_0 denotes the fraction of all sequences that fold into the correct secondary structure, while x_K is the fraction of sterile individuals. With each replication event, the telomere length decreases by 1 if the telomerase is not functional.

With sequence length L, per-nucleotide mutation rate μ and a probability ν that an offspring retains a functional telomerase (the density of the neutral network (Huynen et al., 1996a; Ofria and Adami, 1999; Wilke, 2001; Reidys and Stadler, 2001)), the probability with which a viable sequence gives birth to an offspring that

also resides on the neutral network is

$$Q = [1 - \mu(1 - \nu)]^L \approx \exp^{-L\mu(1-\nu)} \tag{2}$$

Once outside the neutral network, when a sequence X_i, $i \in [1, K-1]$ replicates, it can either become a member of x_{i+1} if it does not recover the correct fold, or become a member of x_0, if it does. More precisely, the replication occurs through (see also Fig. 5)

$$X_0 \xrightarrow{Q} X_0 \tag{3a}$$
$$X_0 \xrightarrow{1-Q} X_1 \tag{3b}$$
$$X_i \xrightarrow{1-p} X_{i+1}, \quad i \in [1, K-1] \tag{3c}$$
$$X_i \xrightarrow{p(1-R)} X_{i+1}, \quad i \in [1, K-1] \tag{3d}$$
$$X_i \xrightarrow{pR} X_0, , \quad i \in [1, K-1] \tag{3e}$$

with the eqs. (3c) and (3d) distinguishing between replication with or without mutation. Here p is the probability of replication with mutation and is defined as

$$p = 1 - (1 - \mu)^L. \tag{4}$$

Notice that we have considered the approximation discussed in the previous section for which R is independent on the telomere length of the sequence, as Fig. 3 justifies. Under the assumptions detailed above, it is now straightforward to derive the temporal evolution of x_i:

$$\begin{aligned}
\dot{x_0} &= pR \sum_{i=1}^{K-1} x_i + Qx_0 - \Phi x_0 \\
\dot{x_1} &= (1 - Q)x_0 - \Phi x_1 \\
\dot{x_i} &= (1 - pR)x_{i-1} - \Phi x_i, \quad i \in [2, K],
\end{aligned}$$

where Φ is a dilution flux that keeps the sum of relative frequencies constant, $\sum_i x_i = 1$. As usual, Φ equals the net production of off-springs. Since the fitness is 1 by definition for all reproducing phenotypes and 0 for the sterile ones, we observe that

$$\Phi = \sum_{i=0}^{K-1} x_i = 1 - x_K$$

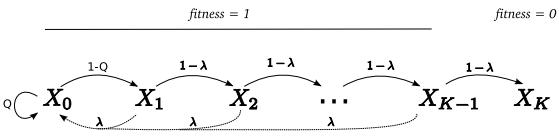

Figure 5: The schematic representation of the model, with $\lambda \equiv pR$ in the equations below. The constant population $N = \sum_{i=0}^{K} N_i$, and $x_i \equiv N_i/N$ characterized by a telomerase length of $(K - i)$.

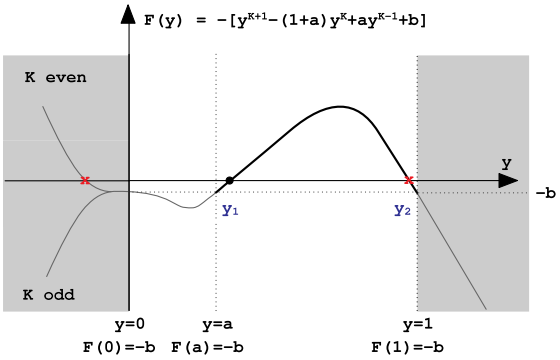

Figure 6: The schematic representation of $F(y)$ defined in eq. (9), with $y \equiv 1 - \bar{x}_K$. This function is defined only for $y \in [a, 1]$, as it results from the condition $\bar{x}_K \geq 0$ in eq. (7). The representation also indicates y_1 to be a fixed point. See below the discussion on y_2. Remember that $y = 0$ implies extinction.

directly measures the fraction of reproducing individuals in the population. Employing the shorthand $q \equiv 1 - Q$, the final form of the equations describing the evolution of the population is thus

$$\dot{x_0} = pR(1 - x_K) - x_0(q + pR - x_K) \tag{5a}$$
$$\dot{x_1} = qx_0 - (1 - x_K)x_1 \tag{5b}$$
$$\dot{x_i} = (1 - pR)x_{i-1} - (1 - x_K)x_i, \quad i \in [2, K] \tag{5c}$$

3.2. Equilibria

The fixed points $\bar{x}_i$ can be expressed in terms of the relative frequency of the $\bar{x}_K$ of the sterile individuals. We have either the trivial solution ($\bar{x}_K = 1$; $\bar{x}_i = 0$, $i < K$), or we obtain, for ($\bar{x}_i > 0, i \leq K$),

$$\bar{x}_0 = \frac{pR(1 - \bar{x}_K)}{q + pR - \bar{x}_K} \tag{6a}$$
$$\bar{x}_1 = \frac{q}{1 - \bar{x}_K}\bar{x}_0 \tag{6b}$$
$$\bar{x}_i = \frac{1 - pR}{1 - \bar{x}_K}\bar{x}_{i-1}, i \in [2, K], \tag{6c}$$

with the last equation providing the condition

$$\bar{x}_K = \frac{qpR}{q + pR - \bar{x}_K}\left(\frac{1 - pR}{1 - \bar{x}_K}\right)^{K-1} \tag{7}$$

This can be rearranged as

$$\bar{x}_K[q + pR - \bar{x}_K](1 - \bar{x}_K)^{K-1} = qpR(1 - pR)^{K-1} \tag{8}$$

Since the r.h.s. is positive for $\mu > 0$, we can immediately conclude that $\bar{x}_K \neq 0$. Moreover, in order to clarify

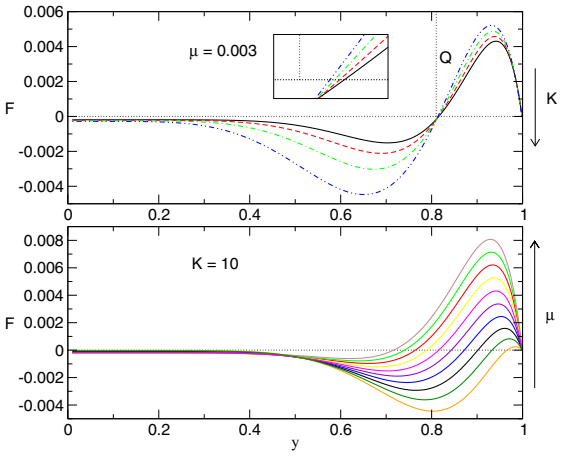

Figure 7: The function $F(y)$ from eq.(9) evaluated using the average values of the recovery fraction R from Fig. 4a. (Upper panel) A fixed value of $\mu = 0.003$ has been chosen, and the function F drawn for $K \in \{7, 8, 9, 10\}$. (Lower panel) A fixed value of $K = 10$, was chosen, and the function was calculated for 10 values in $\mu \in [0.0005 : 0.005]$. As expected, the root y_1 of F is more sensitive to μ than to K.

the solutions of this equation, we shall rewrite it in the variable $y \equiv 1 - \bar{x}_K$:

$$F(y) = -y^{K+1} + (1+a)y^K - ay^{K-1} - b = 0, \quad (9)$$

with new parameters $a \equiv 1-q-pR > 0$ (as it is expected that $Q > pR$) and $b \equiv qpR(1-pR)^{K-1} > 0$. The behavior of this function is sketched in Fig. 6. Notice that the function F, as the system from eq. (5), is valid for $K \geq 2$. For the classical case of non-delayed selection, $K = 1$, the two fixed points are the trivial one (or extinction), $(\bar{x}_0, \bar{x}_1) = (0, 1)$, and the coexistence fixed point, $(\bar{x}_0, \bar{x}_1) = (Q, 1-Q)$, with the former being unstable, and the latter being stable (for μ lower than the extinction threshold). In this $K = 1$ case, the above notation gives the stable fixed point as $y = Q$. Returning to the general case of $K > 1$, the stable fixed point remains in the neighborhood of $y_1 \approx Q$, with y_1 from Fig. 6. Due to the existence of recoveries, one has $y_1 \gtrapprox Q$. For example, for $K = 2$, one has $y_1 = 0.5[Q + \sqrt{Q^2 + 4pR(1-Q)}]$. The second root, y_2 from Fig. 6 does not satisfy the simplex conditions, $\sum \bar{x}_i = 1$, and it is thus not a fixed point of the system in eqs. (5). More precisely, considering $\bar{x}_K = 1 - y_2 \approx 0$, together with the approximation $1 - pR \approx 1$, one has $\sum \bar{x}_i = pR(1+Kq)/(q+pR) < 1$. A numerical verification has been performed too using the values from Fig. 4. For completeness, we remark that for the even-K cases, the function $F(y)$ has another root in the negative quadrant, $y < 0$, which again is not a physically accessible fixed point for our system, as it does not satisfy $\bar{x}_i \geq 0$ for all $i \in [0, K]$.

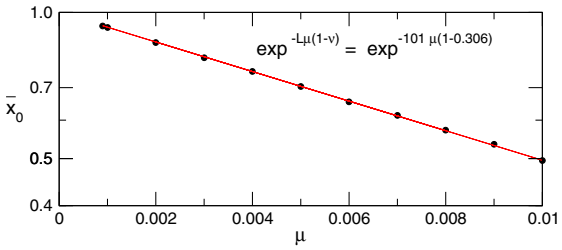

Figure 8: Decay of the steady state concentration, $\bar{x}_0$, as a function of μ for $K = 1$, and the fit to estimate the neutrality ν.

In addition, using estimates of the recovery fraction R from simulations with various parameter settings, we have calculated F numerically. The position of $y_1 \gtrapprox Q$ is nearly independent of K (Fig. 7a), with a pronounced dependence on μ (Fig. 7b). Analytically, for small μ, the Taylor expansion leads to $q \sim L\mu(1-\nu)$ (from eq. 2) and $p \sim L\mu$ (from eq. 4), implying $\bar{x}_K \propto L\mu$. By setting $\bar{x}_k = \xi\mu + O(\mu^2)$ and expanding all quantities in eq.(8) to first order in μ, one gets the following quadratic equation for ξ which indeed does not depend on K:

$$\xi[(1-\nu+R)L - \xi] = (1-\nu)L^2R, \quad (10)$$

The roots of eq. (10) are $\xi_1 = L(1-\nu)$ and $\xi_2 = LR$. It can be seen that ξ_1 can be recovered also from the case $K = 1$ for which $\bar{x}_1 = q \approx L(1-\nu)\mu$ (from eq. 2). For the reproductive but damaged species we observe, by the same arguments employed above, that their dependence on μ, for small μ, also follows $\bar{x}_i \propto \mu$ for all i, again independent of K.

In order to compute $F(y)$ in Fig. 7, the knowledge of Q was required from eq. (2), which in turn relies on an estimate for the neutrality ν. The latter can be determined from the case $K = 1$, considering the approximation $\bar{x}_0 = Q \sim \exp[-L\mu(1-\nu)]$ (Wilke, 2001) (Fig. 8).

The value of $\bar{x}_K$ can be computed numerically using a simple root-finder to solve eq. (9). The values of $\bar{x}_i$, $i \neq K$ are then obtained by rewriting eqs. (6b-6c) as

$$\bar{x}_i = \left(\frac{1-\bar{x}_K}{1-pR}\right)^{K-i} \bar{x}_K, \quad i \in [K-1, 1] \quad (11a)$$

$$\bar{x}_0 = \left(\frac{1-\bar{x}_K}{1-pR}\right)^{K-1} \frac{\bar{x}_K(1-\bar{x}_K)}{q} \quad (11b)$$

This shows that, even though $\bar{x}_K$ is approximately K-independent, as discussed above (Fig. 7 and Fig. 9), the

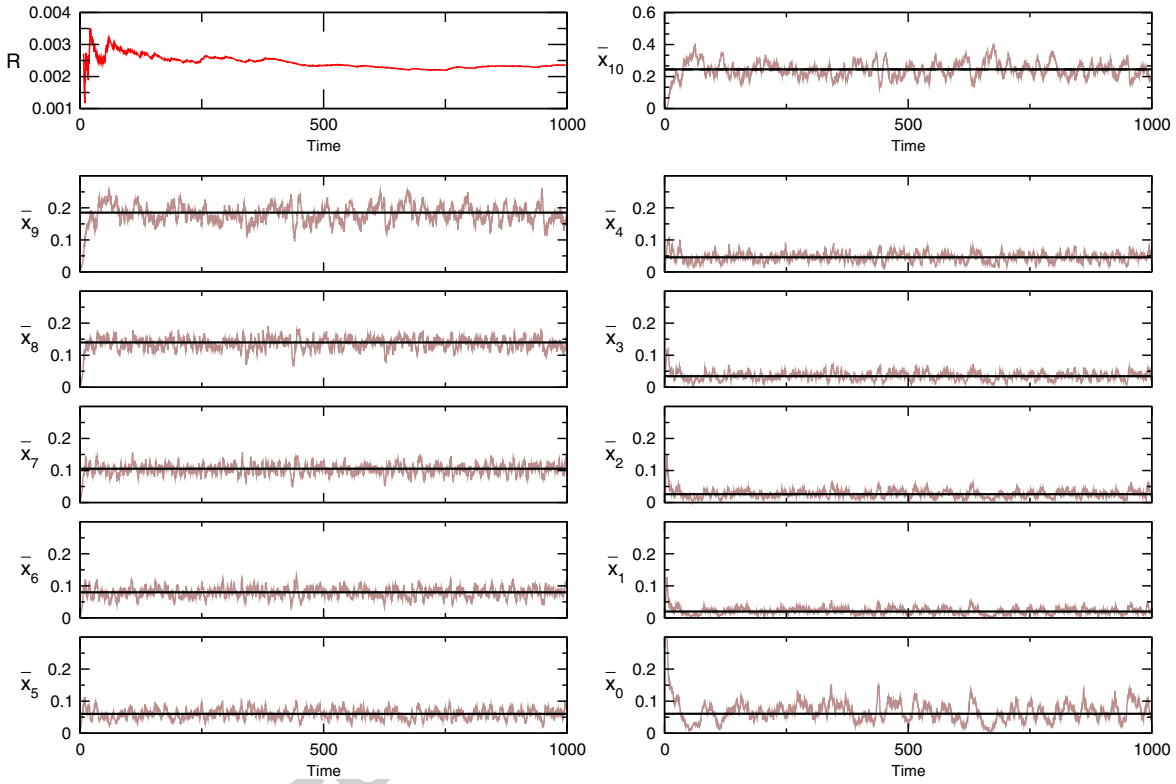

Figure 10: The model from Fig. 5 and the associated eqs. (5) have been superposed on the stochastic experiments. More precisely, (upper left panel) the recovery rate R is estimated from this simulation ($N = 1000, K = 10, \mu = 0.005$) as the ratio of the number of recoveries per mutations off the neutral network (replications with mutation for individuals from x_i with $i \in [1, K-1]$): $R = 0.002359$. Subsequently, the root $y_1 = 0.7549543$ of the function $F(y)$ from eq. (9) is obtained, having calculated $a = 0.754369$ and $b = 0.00019$. In this way, the values of $\bar{x}_i$ are found through eqs. (11).

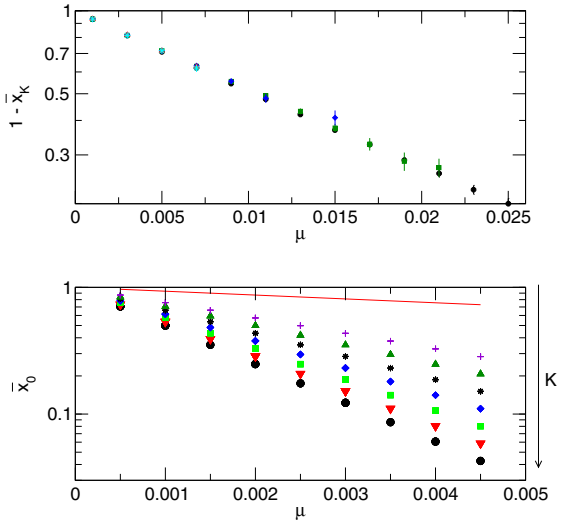

Figure 9: (Upper panel) The average fitness or fraction $\Phi = 1 - x_K$ of reproducing sequences in the population is shown for $K \in \{2, 4, 6, 8\}$. It illustrates that $\bar{x}_K$ is independent of K in very good approximation. (Lower panel) The value of $\bar{x}_0$ estimated through eqs. (11) and based on the recovery probabilities from Fig. 4 with the symbols referring to different values of $K \in [4, 10]$. The straight line results from Fig. 8.

spread of population in the K reproducing-groups is K-dependent. In other words, the unfit ($\bar{x}_K$) and consequently, the fit ($\sum_{i=0}^{K-1} \bar{x}_i = 1 - \bar{x}_K$) levels are independent on K. But the longer the telomere (i.e., the larger the value of K), the wider is the spread or the *dilution* of the population within the fit individuals, owing to the $(K - i)$ exponent. As K increases, $\bar{x}_0$ may reach values dangerously close to 0. We have used the data from Fig. 4 to evaluate $\bar{x}_0$ according to eq. (11b), and in so doing, illustrate the dependence $\bar{x}_0(K, \mu)$. The results are included in the lower panel of Fig. 9. The decrease of $\bar{x}_0$ with K implies that the survival of the species counts exclusively on the probability of recovery. This dilution thus drives extinction at large delays (large K) in a finite population. In other words, in a finite population, delayed selection has the effect of lowering the extinction threshold.

An example of simulation and comparison with the model appears in Fig. 10. Following the reasoning discussed above, the associated values $\bar{x}_i$ from eqs. (11) have been calculated and are shown in Fig. 10 as continuous lines. It can be seen that the mean-field model provides a good fit to the simulations.

3.3. Genetic diversity

It is plausible to assume that lineages with many recoveries in their history and/or recently recovered individuals will preferentially be located at the fringes of the population. Thus they should have a large influence on the sequence evolution. Since the damage/recovery mechanism is capable of bridging gaps in the neutral network, it is tempting to conjecture that this mechanism will also lead to an increased speed of evolution, i.e., an increase in the substitution rate given the same underlying mutation rate μ.

In order to address this issue, we follow the ideas of Huynen et al. (1996b) and investigate the Hamming distance distribution in the population. For each sequence $s \in \mathbb{P}$, let s_j the nucleotide at position j. For each nucleotide $\alpha \in \{A, U, G, C\}$ and $j \in [1, L]$ we consider the fraction

$$\pi_j(\alpha) = \frac{1}{N} \sum_{s \in \mathbb{P}} \delta_{s_j,\alpha} \tag{12}$$

of sequences in the population $\mathbb{P}$ that have nucleotide α at position j. The *profile*, or center of mass, of the population is the $4 \times L$ dimensional vector $\vec{\pi} = \left((\pi_j(\alpha))_{\alpha \in \{A,U,G,C\}}\right)_{j=1}^{L}$. The diversity of the population is conveniently measured by the distribution of pairwise Hamming distances $d_H(s', s'')$, or the distances of the individual sequences to the center of mass. A convenient distance measure is given by the difference between the centroids of the two populations (Derrida and Peliti, 1991; Huynen et al., 1996a; Barnett, 1998). In terms of the profiles, it can be expressed as

$$\Delta^2(\vec{\pi}', \vec{\pi}'') = \sum_{j=1}^{L} \sum_{\alpha \in \{A,U,G,C\}} \left(\pi'_j(\alpha) - \pi''_j(\alpha)\right)^2 \tag{13}$$

Δ^2 therefore directly measures the divergence of the populations. Note any individual sequence s can also be represented by a profile vector $\vec{\pi}^s$ with entries $\pi_j(\alpha) = \delta_{s_j,\alpha}$. In particular, we have

$$\Delta^2(\vec{\pi}^{s'}, \vec{\pi}^{s''}) = 2 d_H(s', s'') \tag{14}$$

as shown in the appendix. The profile distance, eq.(13) thus can also be seen as a straightforward generalization of the Hamming distance.

The speed of evolution can be measured in terms of the mean square displacement of the population over time. More precisely, the motion of the center of mass is captured by the *diffusion constant*

$$D = \lim_{\delta\tau \to 0} \frac{1}{\tau} \left\langle \Delta^2 \left(\vec{\pi}(t + \tau), \vec{\pi}(t)\right) \right\rangle_t \tag{15}$$

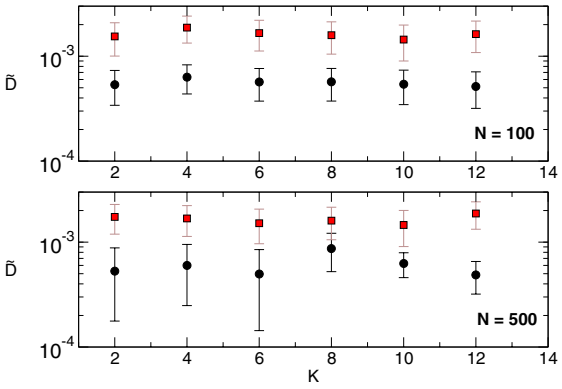

Figure 11: Measuring diffusion through $\tilde{D}$ (see appendix). Ten experiments have been performed for each (N, K) case, and two values of $\mu = 10^{-5}$ (circles) and $\mu = 3 \times 10^{-5}$ (squares). Two tank capacities are used for verification: $N = 100$ (upper panel) and $N = 500$ (lower panel). The diffusion was measured as $\tilde{D} = \langle D \rangle \pm \sqrt{\frac{1}{n} \frac{\sum (D_i - \langle D \rangle)^2}{n-1}}$, with D_i, the mean value of the i^{th} experiment.

where $\langle \, . \, \rangle_t$ denotes the average over time t and simulation runs. The diffusion constant D is a convenient way of estimating the substitution rate directly from simulated populations (Huynen et al., 1996a; Stadler, 2002; Stephan-Otto Attolini and Stadler, 2006), which is independent of the particular rules of the selection/mutation process. It corresponds to the substitution rate used in phylogenetic analysis.

Note that the definitions of D above depends on the ability to explicitly compute the center of mass $\vec{\pi}$ of a population $\mathbb{P}$. Conceptually this means that we need to be able to treat the individual members $s \in \mathbb{P}$ as vectors. This is straightforward in the absence of insertions and deletions because the Hamming distance already is of the appropriate form, see eq.(14). In the presence of insertions and deletions, however, it seems non-trivial to find a vector-space representation for the Levenshtein distance. An alternative vector-based distance measure could be obtained by first constructing an alignment of all sequences in the two populations and then treating gaps as an additional letter in eq.(12). The need to compare populations at different times in eq.(15) complicates the issue, requiring at least alignments of pairs of populations. In contrast, a distance measure between the individuals of the population is sufficient to quantify e.g. the diversity in the population. This begs the question whether eq.(15) can be generalized to a more general setting. In the appendix we show how this can be achieved by introducing an equivalent measure of diffusion through $\tilde{D}$.

Finally, to answer the question whether the delayed selection could leave a fingerprint in the population diversity, we have measured the diffusion coefficient $\tilde{D}$ for various cases of mutation rate. First, we do not expect to see any consequence of the delayed selection when the mutation rate is low enough to impede recoveries. Without recoveries, the enrichment of the population does not occur. We have verified this statement by performing simulations for two values of mutation rate and population number (Fig. 11). Formally, we expect the recoveries to be $R \propto Kp^2 \approx KL^2\mu^2$, where the exponent in the latter statement refers to the two mutations needed to fall off and return to the neutral network, respectively. For the μ values employed in the examples from Fig. 11, recoveries are negligible and no differential diffusion is observed for different time delays. Just for verification, the population number does not affect the diffusion coefficient.

For mutation rates that allow a significant number of recoveries, we expect to see that recoveries lead to a higher evolutionary rate identified by a higher diversity, and thus a higher diffusion coefficient. The diffusion coefficient $\tilde{D}$ was measured for: only the replicating (viable) sequences (l.h.s panel of Fig. 12), and only the undamaged sequences (on the neutral network) (r.h.s panel of Fig. 12). Even though no increased diffusion is observed for the viable individuals, a slight effect can be observed for the neutral network. It illustrates the expectation that, at significant mutation rates (high, but not too high; see Fig. 4), a more efficient exploration of the neutral network is provided by the delayed selection. When measured at the level of the entire viable population (l.h.s panel), this slight increase is lost in the accumulation of damaged individuals.

These results show that the delayed selection leaves no unequivocal traces of its presence in terms of population diversity. We conclude therefore that even massive delays in the effect of selection do not appreciably affect substitution rates.

4. Discussion

The scenario of classical Darwinian selection considers that selection instantaneously punishes or rewards changes in fitness associated to the individual genomes. This scenario is employed in the overwhelming majority of studies regarding fitness-driven selection. Only a few studies considered fitness effects that reach across generations. For instance, Wilke (2002) considers a model in which fitness is the product of a maternal contribution and the offsprings own genotype. Still, selection acts instantaneous to remove lethal genotypes

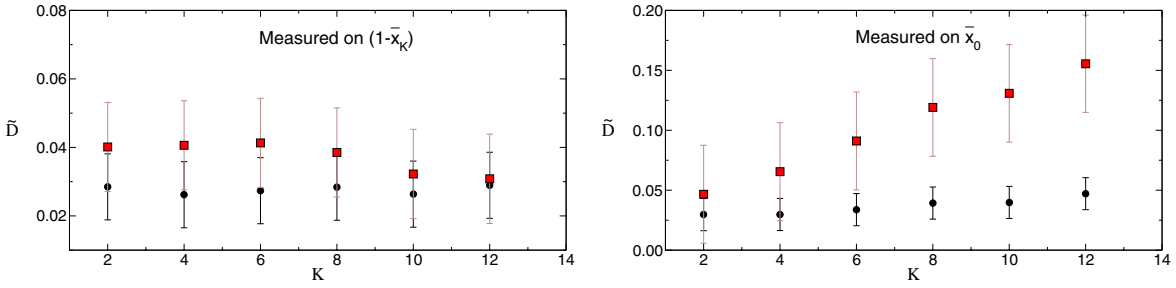

Figure 12: Measuring diffusion through $\tilde{D}$. As in Fig.11, ten experiments have been performed for each K-case (here $N = 1000$), and two values of $\mu = 10^{-3}$ (circles) and $\mu = 3 \times 10^{-3}$ (squares). (l.h.s panel) The diffusion was measured only from the replicating (viable) sequences. (r.h.s panel) From the same experiments in the lhs panel, only the sequences of maximum telomere length (sequences on the neutral network) were extracted and used for measuring the diffusion.

from the population. In our model, which is inspired by eukaryotic telomere damage, individuals carrying dysfunctional genomes are oblivious of this fact for several generations. At least intuitively, this setup should emphasize the effects of delayed selection as much as possible compared to more realistic scenarios in which genetic damage is associated also with some instantaneous fitness effects.

The model is investigated in two settings: stochastic computer simulations based on neutral networks of RNA secondary structures, and a deterministic infinite-population-size model. The RNA-based simulations show that damage-and-recovery is a frequent phenomenon for a wide range of mutation rate values. In particular, after a relatively short time, all individuals in the population derive from ancestors that have sustained damage and have subsequently recovered through compensatory mutations. We have demonstrated, furthermore, that it is sufficient to estimate a few parameters, namely the recovery rates R and the degree of neutrality ν to parameterize the deterministic ODE model in such a way that it reproduces the phenomena observed in the stochastic simulation.

For simplicity, the deterministic system was set up as a flow reactor under constant organization like Eigen's Quasispecies Model. In this system, we observe a single stable equilibrium in which x_0, the fraction of undamaged individuals, and x_K, the fraction of sterile members, strike a balance that depends primarily on the mutation rate μ. For large values of μ, x_0 becomes very small and thus fluctuations can easily wipe out the undamaged part of the population. This behavior roughly corresponds to the extinction threshold. Therefore, the main effect of delayed selection is to reduce the critical mutation rate. In other words, as one may have expected, genetic components evolving under delayed selection have an increased risk of being lost.

To our surprise, however, delayed selection does not appear to have a measurable effect on the substitution rates observed at population level. Delayed selection, therefore does not easily reveal itself in genomic DNA sequences. For one, this begs the question whether there are more subtle effects on substitution rates. If they exist, they will presumably depend on the specifics of the selection pressures of the particular protein or RNA in question. On the other hand, the apparently small impact of delayed selection at the sequence level could hide that this is indeed a rather frequent phenomenon. The well-known observation that deletion of a highly conserved gene often has no appreciable phenotype at least under laboratory conditions could be related to our topic.

In conclusion, we have approached a question that has not been addressed so far in neither modeling nor simulating framework, by studying the scenario in which the selection of the fittest is delayed for several generations. Such a scenario occurs in very unrelated topics, from telomere damage-repair system to plant breeding. By this study we have thus laid the grounds of further explorations of the consequences of such a scenario.

Acknowledgements

This work was supported by the EU 6th Framework project SYNLET (NEST 043312) and AGAUR - Departament d'Universitats, Recerca i Societat de la Informacion de la Generalitat de Catalunya. AM thanks Bernat Corominas-Murtra for useful insights.

References

Babajide, A., Farber, R., Hofacker, I. L., Inman, J., Lapedes, A. S., Stadler, P. F., Sep 2001. Exploring protein sequence space using knowledge-based potentials. J Theor Biol 212 (1), 35–46.

Babajide, A., Hofacker, I. L., Sippl, M. J., Stadler, P. F., 1997. Neutral networks in protein space: a computational study based on knowledge-based potentials of mean force. Fold Des 2 (5), 261–269.

Barnett, L., 1998. Ruggedness and neutrality - the nkp family of fitness landscapes. In: Adami, C., Belew, R. K., Kitano, H., Taylor, C. (Eds.), Proceedings of the sixth international conference on Artificial life. MIT Press, pp. 18–27.

Bastolla, U., Roman, H. E., Vendruscolo, M., Sep 1999. Neutral evolution of model proteins: diffusion in sequence space and overdispersion. J Theor Biol 200 (1), 49–64.

Bull, J. J., Sanjun, R., Wilke, C. O., Mar 2007. Theory of lethal mutagenesis for viruses. J Virol 81 (6), 2930–2939.

Derrida, B., Peliti, L., 1991. Evolution in a flat fitness landscape. Bull. Math. Biol. 53, 355–382.

Eigen, M., 1971. Selforganization of matter and the evolution of biological macromolecules. Die Naturwissenschaften 10, 465–523.

Eigen, M., McCaskill, J., Schuster, P., 1989. The molecular Quasispecies. Adv. Chem. Phys. 75, 149 – 263.

Fontana, W., Schnabl, W., Schuster, P., 1989. Physical aspects of evolutionary optimization and adaption. Phys. Rev. A 40, 3301–3321.

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

Forst, C. V., Reidys, C., Weber, J., 1995. Evolutionary dynamics and optimization: neutral networks as model-landscape for RNA secondary structure folding landscapes. In: Morán, F., Moreno, A., Merelo, J. J., Chacón, P. (Eds.), *Advances in Artificial Life,* Lecture Notes in Artificial Intelligence **929**. Springer, Berlin, pp. 128–147.

Gruener, W., Giegerich, R., Strothmann, D., Reidys, C., Weber, J., Hofacker, I. L., Stadler, P. F., Schuster, P., 1996. Analysis of RNA sequence structure maps by exhaustive enumeration. I. neutral networks. Monath. Chem. 127, 355–374.

Hartman IV, J. L., Garvik, B., Hartwell, L., 2001. Principles for the buffering of genetic variation. Science 291, 1001–1004.

Herrera, E., Samper, E., Martín-Caballero, J., Flores, J. M., Lee, H. W., Blasco, M. A., 1999. Disease states associated with telomerase deficiency appear earlier in mice with short telomeres. EMBO J. 18, 2950–2960.

Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M., Schuster, P., 1994. Fast folding and comparison of RNA secondary structures. Monatsh. Chem. 125, 167–188.

Huynen, M. A., Stadler, P. F., Fontana, W., 1996a. Smoothness within ruggedness: the role of neutrality in adaptation. Proc. Natl. Acad. Sci. (USA) 93, 397–401.

Huynen, M. A., Stadler, P. F., Fontana, W., Jan 1996b. Smoothness within ruggedness: the role of neutrality in adaptation. Proc Natl Acad Sci U S A 93 (1), 397–401.

Jorg, T., Martin, O. C., Wagner, A., 2008. Neutral network sizes of biological RNA molecules can be computed and are not atypically small. BMC Bioinformatics 9, 464.

Kaelin Jr, W. G., 2005. The concept of synthetic lethality in the context of anticancer therapy. Nat Rev Cancer 5, 689–698.

Le Meur, N., Gentleman, R., 2008. Modeling synthetic lethality. Genome Biol 9, R135.

Lee, H.-W., Blasco, M. A., Gottlieb, G. J., Horner II, J. W., Greider, C. W., DePinho, R. A., 1998. Essential role of mouse telomerase in highly proliferative organs. Nature 392, 569–574.

Niida, H., Matsumoto, T., Satoh, H., Shiwa, M., Tokutake, Y., Furuichi, Y., Shinkai, Y., 1998. Severe growth defect in mouse cells lacking the telomerase RNA component. Nat. Genet. 19, 203–206.

Ofria, C. A., Adami, C., 1999. Evolution of genetic organization in digital organisms. In: Landweber, L., Winfree, E. (Eds.), Evolution as Computation. Springer, pp. 167–182.

Reidys, C., Forst, C. V., Schuster, P., Jan 2001. Replication and mutation on neutral networks. Bull Math Biol 63 (1), 57–94.

Reidys, C., Stadler, P. F., Schuster, P., Mar 1997. Generic properties of combinatory maps: neutral networks of rna secondary structures. Bull Math Biol 59 (2), 339–397.

Reidys, C. M., Stadler, P. F., 2001. Neutrality in fitness landscapes. Appl. Math. & Comput. 117, 321–350.

Schuster, P., Fontana, W., Stadler, P. F., Hofacker, I. L., 1994. From sequences to shapes and back: A case study in RNA secondary structures. Proc. Roy. Soc. Lond. B 255, 279–284.

Stadler, B. M., Stadler, P. F., Wagner, G. P., Fontana, W., Nov 2001. The topology of the possible: formal spaces underlying patterns of evolutionary change. J Theor Biol 213 (2), 241–274.

Stadler, B. M. R., 2002. Diffusion of a population of interacting replicators in sequence space. Adv. Complex Systems 5 (4), 457–461.

Stephan-Otto Attolini, C., Stadler, P. F., 2006. Evolving towards the hypercycle: A spatial model of molecular evolution. Physica D 217, 134–141.

Sumedha, Martin, O. C., Wagner, A., 2007. New structural variation in evolutionary searches of RNA neutral networks. Biosystems 90, 465–485.

Urquhart, F. A., 1960. The monarch butterfly. University of Toronto Press, Toronto.

Verdun, R. E., Karlseder, J., 2007. Replication and protection of telomeres. Nature 447, 924–931.

Wilke, C. O., 2001. Selection for fitness versus selection for robustness in RNA secondary structure folding. Evolution 55, 2412–2420.

Wilke, C. O., Feb 2002. Maternal effects in molecular evolution. Phys Rev Lett 88 (7), 078101.

Xie, M., Mosig, A., Qi, X., Li, Y., Stadler, P. F., Chen, J. J.-L., 2008. Size variation and structural conservation of vertebrate telomerase RNA. J. Biol. Chem. 283, 2049–2059.

Zhu, H., Casselman, A., Reppert, S. M., 2008. Chasing migration genes: A brain Expressed Sequence Tag resource for summer and migratory monarch butterflies (*Danaus plexippus*). PLoS ONE 3, e1345.

Appendix

Measuring Diffusion in Metric Spaces

Let $A = \{\vec{x_1}, \ldots, \vec{x_{n_A}}\}$ and $B = \{\vec{y_1}, \ldots, \vec{y_{n_B}}\}$ be two finite sets of vectors in some vector space $\mathbb{V}$. As an example, in the present case, the sequence *AACGT* can be written in the base $\{A, G, C, T\}$ as 1000 1000 0010 0100 0001.

a vector in Our goal is to express the mean square displacement

$$\Delta^2 = \Delta^2(A, B) = \left(\frac{1}{n_A} \sum_{i \in A} \vec{x_i} - \frac{1}{n_B} \sum_{i \in B} \vec{y_i} \right)^2 \qquad (16)$$

of the centers of gravity of A and B in terms of distances between their elements. In a Euclidean vector space, we

have canonical distances given by $d_{ij}^2 = (\vec{x}_i - \vec{x}_j)^2$ for $i, j \in A$, $d_{ij}^2 = (\vec{y}_i - \vec{y}_j)^2$ for $i, j \in B$, and $d_{ij}^2 = (\vec{x}_i - \vec{y}_j)^2$ for $i \in A$ and $j \in B$. It is convenient to introduce the following quantities, which can be computed in terms of pairwise distances:

$$\begin{aligned} V_A &= \frac{1}{n_A^2} \sum_{i\in A} \sum_{j\in A} (\vec{x}_i - \vec{x}_j)^2 = \frac{1}{n_A^2} \sum_{i\in A} \sum_{j\in A} d_{ij}^2 \\ V_B &= \frac{1}{n_B^2} \sum_{i\in B} \sum_{j\in B} (\vec{y}_i - \vec{y}_j)^2 = \frac{1}{n_B^2} \sum_{i\in B} \sum_{j\in B} d_{ij}^2 \\ W &= \frac{1}{n_A n_B} \sum_{i\in A} \sum_{j\in B} (\vec{x}_i - \vec{y}_j)^2 = \frac{1}{n_A n_B} \sum_{i\in A} \sum_{j\in B} d_{ij}^2 \end{aligned} \tag{17}$$

In the following we will prove the identity

$$\Delta^2 = W - V_A/2 - V_B/2 \tag{18}$$

which shows that the mean square displacement can be expressed in terms of pairwise distances.

First observe that $\left(\sum_{i\in A} \vec{x}_i\right)^2 = \sum_{i,j\in A} \vec{x}_i \vec{x}_j$ and hence

$$n_A^2 V_A = 2n_A \sum_{i\in A} \vec{x}_i^{\,2} - 2 \sum_{i,j\in A} \vec{x}_i \vec{x}_j = 2n_A \sum_{i\in A} \vec{x}_i^{\,2} - 2\left(\sum_{i\in A} \vec{x}_i\right)^2$$

An analogous expression holds for V_B. Next we rewrite the definition of Δ^2 in the form

$$\Delta^2 = \frac{1}{n_A^2 n_B^2} \left(n_B^2 \left(\sum_{i\in A} \vec{x}_i\right)^2 + n_A^2 \left(\sum_{i\in B} \vec{y}_i\right)^2 - 2n_A n_B \sum_{i\in A} \sum_{j\in B} \vec{x}_i \vec{y}_j \right)$$

and use this expression to compute

$$\begin{aligned} & n_A^2 n_B^2 W - n_A^2 n_B^2 \Delta^2 \\ &= n_A n_B^2 \sum_{i\in A} \vec{x}_i^{\,2} + n_B n_A^2 \sum_{j\in B} \vec{y}_j^{\,2} - n_B^2 \left(\sum_{i\in A} \vec{x}_i\right)^2 - n_A^2 \left(\sum_{i\in B} \vec{y}_i\right)^2 \\ &= n_B^2 \left(n_A \sum_{i\in A} \vec{x}_i^{\,2} - \left(\sum_{i\in A} \vec{x}_i\right)^2 \right) + n_A^2 \left(n_B \sum_{j\in B} \vec{y}_j^{\,2} - \left(\sum_{i\in B} \vec{y}_i\right)^2 \right) \\ &= \frac{n_B^2}{2} n_A^2 V_A + \frac{n_A^2}{2} n_B^2 V_B \end{aligned}$$

Eq.(18) now follows immediately.

Returning to the definition of RNA sequences as vectors, eq. (16) coincides with eq. (13) which employs a different notation for the population profiles. And thus, the distance between two vectors as included in eq. (17) can be written as

$$\begin{aligned} d^2(\vec{x}, \vec{y}) &= \sum_{j=1}^{n} \sum_{\alpha\in\{A,U,G,C\}} \left(x_{j,\alpha} - y_{j,\alpha}\right)^2 \\ &= 2\, d_H(\vec{x}, \vec{y}) \end{aligned}$$

where $d_H(\vec{x}, \vec{y})$ is the Hamming distance between the two sequences.

The importance of eq.(18) is twofold. First, it implies that the diffusion coefficient

$$\tilde{D} \equiv \lim_{\tau\to 0} \frac{\Delta^2(A_{t+\tau}, A_t)}{\tau} \tag{19}$$

is a metric quantity at heart that does not necessarily require the explicit computation of the "centers of gravity" of the populations at the different time points. Secondly, it suggests eq. (18) to be the *definition* of Δ^2 in situations where $\mathbb{V}$ is not given explicitly, or where we only have a metric structure at our disposal. Eq.(18) thus is of practical use, since pairwise distances of sequences in related populations can be computed efficiently, while the construction of good multiple sequence alignments may be quite tedious.