Evolution dynamics of a model for gene duplication under adaptive conflict

Mark Ancliff and Jeong-Man Park[*]

Department of Physics, The Catholic University of Korea, Bucheon 420-743, Korea

(Received 28 January 2014; published 3 June 2014)

We present and solve the dynamics of a model for gene duplication showing escape from adaptive conflict. We use a Crow-Kimura quasispecies model of evolution where the fitness landscape is a function of Hamming distances from two reference sequences, which are assumed to optimize two different gene functions, to describe the dynamics of a mixed population of individuals with single and double copies of a pleiotropic gene. The evolution equations are solved through a spin coherent state path integral, and we find two phases: one is an escape from an adaptive conflict phase, where each copy of a duplicated gene evolves toward subfunctionalization, and the other is a duplication loss of function phase, where one copy maintains its pleiotropic form and the other copy undergoes neutral mutation. The phase is determined by a competition between the fitness benefits of subfunctionalization and the greater mutational load associated with maintaining two gene copies. In the escape phase, we find a dynamics of an initial population of single gene sequences only which escape adaptive conflict through gene duplication and find that there are two time regimes: until a time t^* single gene sequences dominate, and after t^* double gene sequences outgrow single gene sequences. The time t^* is identified as the time necessary for subfunctionalization to evolve and spread throughout the double gene sequences, and we show that there is an optimum mutation rate which minimizes this time scale.

DOI: 10.1103/PhysRevE.89.062702 PACS number(s): 87.23.Kg, 87.10.−e, 64.10.+h, 03.50.−z

I. INTRODUCTION

Life inevitably depends on protein function. Proteins constitute the phenotypes which result from the expression of an organism's genes as well as the influence of environmental factors and are the level at which natural selection acts. To understand how evolution works it is therefore essential to know how proteins evolve, because in a constantly changing environment proteins with new functions determine how successfully an organism can survive and reproduce. It is accepted in general that new proteins and new protein functions evolve from existing ones, either through small-scale mutations such as point mutations or through large-scale mutations such as recombinations and gene duplications. Although the mechanism of the emergence and evolution of new proteins and new protein functions is still unresolved, gene duplication is known to play an important role [1]. Recent researches lead to the predominant view that a gene duplication is required for biological innovation because it provides the opportunity for evolution to try out alternative protein designs without sacrificing an existing design [2–4].

Gene duplication is the process by which a chromosome or a portion of DNA that contains a gene is duplicated. Gene duplications can arise as products of several types of errors in DNA replication and repair machinery. Small-scale duplications of one or a few genes can happen by unequal crossing over or retrotransposition. Unequal crossing over can occur when two chromosomes cross over and recombine. During recombination, strands of two chromosomes break and rejoin to the opposite chromosome, so that genetic information is moved from one chromosome to another. If the recombination is unequal, duplication can result. Retrotransposition is the process in which a sequence of DNA is copied to RNA and then copied back to DNA instead of being translated into proteins. This results in extra copy of the same sequence of DNA and any genes found along this sequence will be duplicated in the process. Whole-genome duplications (WGDs) are another possible mechanism, in which the entire chromosome is replicated twice. WGDs are the result of mitotic cell divisions that duplicate the genome but fail to separate the copied genome from the original.

After gene duplication, fates of duplicated genes are still under much debate. There are several models about the evolutionary mechanisms that are responsible for the retention and subsequent divergence of newly created gene duplicates [5–7]. Among them, two dominant models are the neofunctionalization (NEO-F) model, when an original gene has one dominant function [8], and the escape from adaptive conflict (EAC) model, when an original gene has two or more distinct subfunctions [9,10]. According to the NEO-F model, after duplication due to the functional redundancy, degenerative mutations are accumulated in one copy, while the other duplicate copy continues to perform the essential tasks of the ancestral single-copy gene. In the majority of cases, the redundant gene duplicate will eventually be rendered functionless by accumulated inactivating mutations and becomes a pseudogene, a gene that is no longer transcribed. In a very small minority of cases, the redundant gene may escape this fate by fixing one or more mutations that fortuitously adapt the encoded protein to a new function.

When an ancestral single-copy gene encodes a generalist protein that is capable of performing two or more distinct subfunctions, the product of this gene experiences "adaptive conflict," as joint optimization of the protein's multiple subfunctions is constrained by antagonistic pleiotropy. On a single-copy pleiotropic gene, mutations that increase activity of one function may be prevented from going to fixation because they reduce activity of the other function and vice versa. If this gene is duplicated, then each of the two duplicates

[*] jmanpark@catholic.ac.kr

can break free of these pleiotropic constraints and specialize on activity of one function, respectively. In the EAC model, gene duplication can resolve adaptive conflicts between competing subfunctions of a pleiotropically constrained single-copy gene by the division of labor between the duplicated genes brought about by the fixation of advantageous mutations that refine ancestral subfunctions of the encoded protein [11].

In this paper, we present a statistical physics method to analyze the EAC model for gene duplication. We use the Crow-Kimura quasispecies model [also known as a parallel mutation-selection (ParaMuSe) model] of evolution to describe the dynamics of a mixed population of individuals with single and double copies of a pleiotropic gene with two functions. The asymptotic mean fitness of the population in this model was analyzed in an earlier paper [12]. Here we solve the population dynamics toward the asymptotic state. By mapping the evolutionary dynamics onto the dynamics of a quantum spin chain, we derive the spin coherent state path integral representation and solve the evolutionary dynamics under a saddle-point semiclassical approximation. The model shows a phase transition between an EAC phase in which each copy of a duplicated gene evolves toward subfunctionalization and a duplication loss of function (DLoF) phase in which one copy maintains its pleiotropic function and the other copy undergoes neutral mutation to lose its original function. We analyze the population dynamics toward the EAC phase and calculate the mean time taken for the duplicated gene to become fixed in the population.

The paper is organized as follows. In Sec. II, we introduce the Crow-Kimura quasispecies model of evolution and map the Crow-Kimura model onto a quantum spin model with a Hamiltonian composed of Pauli spin operators. By using the spin coherent state path integral representation, we show how to solve the dynamics of the Crow-Kimura model semiclassically for the general fitness landscape in Sec. III. In Sec. IV, we extend the mapping and the path integral representation to the EAC model of gene duplication and derive the phase diagrams and evolutionary dynamics toward the EAC phase. In Sec. V, we conclude with a brief discussion and outline possible future works.

II. MAPPING OF QUASISPECIES MODELS TO QUANTUM SYSTEMS

We first introduce the ParaMuSe model [13], which describes the evolution of an infinite population of sequences subject to mutation and selection. In this quasispecies model, the sequences can be written as a chain of L spins, $S^i = (s^i_1,\dots,s^i_L)$, $i = 1,2,\dots,2^L$, where $s^i_j \in \{\uparrow,\downarrow\}$, $j = 1,2,\dots,L$. Each sequence represents a different genotype, and each spin in a sequence represents a base pair in the genome. For sequences of length L, this gives a total of 2^L possible genotypes. The population is defined as a distribution on the set of sequences, $p(S^i)$.

The (linearized) ParaMuSe model describes the evolution of the population $\{p(S^i)\}$ by the equation

$$\frac{dp(S^i)}{dt} = f(S^i)p(S^i) + \mu \sum_{d(i,j)=1} [p(S^j) - p(S^i)], \tag{1}$$

where the sum runs over sequences S^j a single spin flip away from sequence S^i [$d(i,j) = 1$]. [The Hamming distance between S^i and S^j, $d(i,j)$, is defined as the number of spin sites at which the two sequences differ.] The function $f(S^i)$ is known as the *fitness* and represents the reproduction rate of the sequence S^i. The second term on the right describes mutation of the sequence S^j into sequence S^i (and vice versa), where μ is the mutation rate.

For a given fitness function, the ParaMuSe model has unique asymptotic population distributions defined by

$$p_{\text{asp}}(S^i) := \lim_{t\to\infty} \frac{p(S^i)}{\sum_j p(S^j)}. \tag{2}$$

The asymptotic population distribution has a population growth rate

$$\overline{f} = \sum_j p_{\text{asp}}(S^j) f(S^j) \tag{3}$$

known as the *asymptotic mean fitness* of the population.

We now map this model onto a quantum spin model with a Hamiltonian composed of Pauli spin operators. Let $V_{1/2}$ denote the spin-$\frac{1}{2}$ Hilbert space

$$V_{1/2} = \{a|\uparrow\rangle + b|\downarrow\rangle\}; \tag{4}$$

then, the population distribution $p(S^i)$ can be considered as a state, $|\Psi_p\rangle$, in the space $\otimes^L V_{1/2}$:

$$|\Psi_p\rangle = \sum_i p(S^i)||S^i\rangle\rangle. \tag{5}$$

The ParaMuSe model dynamics can be expressed by an imaginary-time Hamiltonian operator on this Hilbert space [14–16]. We have

$$\frac{d}{dt}|\Psi(t)\rangle = -H|\Psi(t)\rangle, \tag{6}$$

with

$$H = -f(\sigma^z_1,\dots,\sigma^z_L) - \mu \sum_{i=1}^{L} (\sigma^x_i - 1), \tag{7}$$

where the σ_i are the Pauli spin operators acting on the ith spin in the sequence. The ground states of the Hamiltonian H correspond to the asymptotic population distributions of the ParaMuSe model.

In general, the fitness landscape $f(S^i)$ is an arbitrary function of the configuration of the sequence. For simplicity, we consider the case in which the fitness $f(S^i)$ depends only on the sum of spins in the sequence S^i; i.e., the fitness is a function of one parameter, the Hamming distance from the reference sequence $S^1 = (\uparrow\uparrow\cdots\uparrow)$:

$$f(\sigma^z_1,\dots,\sigma^z_L) = f(\sigma^z_1 + \cdots + \sigma^z_L). \tag{8}$$

We refer to such a fitness landscape as a symmetric fitness landscape because it has a permutation symmetry (later on we will generalize to the case where fitness is a function of the Hamming distance from several reference sequences [17]). In this case, we can rewrite the ParaMuSe Hamiltonian in terms of the total spin operators, $\sigma^\alpha = \sigma^\alpha_1 + \cdots + \sigma^\alpha_L$ ($\alpha = x,y,z$).

The ParaMuSe Hamiltonian becomes

$$H = -f(\sigma^z) - \mu(\sigma^x - L). \tag{9}$$

Since the Hamiltonian involves only total spin operators, we can decompose the Hilbert space $\otimes^L V_{1/2}$ into a sum of irreducible subspaces (representations) under the action of the total spin operators $\{J_x, J_y, J_z\}$, $J_\alpha = \frac{1}{2}\sigma^\alpha$ and consider the action of the Hamiltonian on each subspace separately. The decomposition is as follows:

$$\otimes^L V_{1/2} = V_{L/2} \oplus \left(\oplus_{k=1}^{[L/2]} c_k V_{L/2-k}\right), \tag{10}$$

where V_j is the spin-j representation and $[x]$ denotes the integer part of x. The coefficients c_k are not important because we will focus on the highest-spin subspace $V_{L/2}$. The highest-spin subspace $V_{L/2}$ in the decomposition above can be identified as the subspace of all populations which are invariant under permutations in the order of spins [12]. From the Perron-Frobenius theorem, it is known that the ParaMuSe Hamiltonian has a unique ground state, which corresponds to the asymptotic population distribution of Eq. (1). The ground state for a symmetric fitness landscape must be invariant under permutations (if it were not, then we could permute the spins to create another ground state, violating uniqueness), and so it belongs to the spin-$L/2$ subspace, $V_{L/2}$. For large L, we apply the techniques of the spin coherent state path integral to the $V_{L/2}$ space to extract expressions for the ground state and the ground state energy.

III. SPIN COHERENT STATE PATH INTEGRAL

The dynamics of the ParaMuSe model in the case of symmetric populations and fitness functions corresponds to the imaginary-time dynamics of a quantum spin-$L/2$ state under the action of a Hamiltonian composed of Pauli spin operators [12]. A path integral for the quantum system can be constructed using spin coherent states [18,19].

In the spin-j space with $j = L/2$, there are $(2j+1)$ number of normalized orthogonal basis vectors $|j,m\rangle$, $m = -j, -j+1, \ldots, j$, defined as common eigenvectors of J^2 and J_z:

$$J^2\,|j,m\rangle = j(j+1)\,|j,m\rangle\,, \tag{11}$$

$$J_z\,|j,m\rangle = m\,|j,m\rangle\,. \tag{12}$$

We define un-normalized orthogonal vectors $|j,m)$ as

$$\begin{aligned} |j,m) &= \frac{(j-m)!}{(2j)!} J_+^{j+m}\,|j,-j\rangle \\ &= \sqrt{\frac{(j+m)!(j-m)!}{(2j)!}}\,|j,m\rangle\,, \end{aligned} \tag{13}$$

with $(j,m|j,m) = d!(L-d)!/L!$ so that

$$J_+\,|j,m) = (j-m)\,|j,m+1)\,, \tag{14}$$

$$J_-\,|j,m) = (j+m)\,|j,m-1)\,, \tag{15}$$

where $J_\pm = J_x \pm i J_y$, and identify these vectors as the population states $|d) = |j,m)$ with $d = j+m$, $j = L/2$, and $d = 0,1,2,\ldots,L$ where d is a Hamming distance of a state from the state $|d=0) = |\downarrow,\downarrow,\cdots,\downarrow\rangle$. With a definition of the projection state

$$(\cdot| = (0|\,e^{J_-}, \tag{16}$$

we find

$$(\cdot|d) = 1, \tag{17}$$

and we write a normalized state $|\Psi(t))$

$$|\Psi(t)) = \sum_{d=0}^{L} p(d;t)\,|d)\,, \tag{18}$$

so that $(\cdot|\Psi(t)) = 1$ for all $t \geqslant 0$ and $p(d;t) = C_d^L\,(d|\Psi(t))$ with $C_d^L = L!/(L-d)!d!$.

Now we define a family of spin coherent states $|z\rangle$ by

$$|z\rangle = \exp[zJ_+]\,|0) = \sum_{d=0}^{L} \frac{L!}{(L-d)!d!} z^d\,|d)\,, \tag{19}$$

where z is a complex number, $|0)$ denotes the lowest eigenvalue state with respect to J_z normalized so that $(0|0) = 1$, and $|z=1\rangle = |\cdot)$ with $(\cdot|z\rangle = (0|\,e^{J_-}e^{zJ_+}\,|0) = (1+z)^L$. By making use of the resolution of the identity

$$I = \frac{L+1}{2\pi}\iint \frac{d\bar{z}dz}{(1+\bar{z}z)^{N+2}}|z\rangle\langle z|, \tag{20}$$

the population state $|d)$ can be expressed in terms of spin coherent states

$$|d) = \frac{L+1}{2\pi}\iint \frac{d\bar{z}dz}{(1+\bar{z}z)^{L+2}}\bar{z}^d|z\rangle, \tag{21}$$

with overlaps $(d|z\rangle = z^d$ and $\langle z|d) = \bar{z}^d$. In the quasispecies framework $|0)$ denotes the population where the sequence $(\downarrow\downarrow\cdots\downarrow)$ has frequency 1 and all other sequences are unpopulated and $|\cdot) = |z=1\rangle$ corresponds to a uniformly distributed population in which each sequence hasxbrk frequency 1.

In Ref. [12] we gave a different definition of spin coherent states, where $|\theta,\phi\rangle$ denoted the unique state such that

$$(\sin\theta\cos\phi\,\sigma^x + \sin\theta\sin\phi\,\sigma^y + \cos\theta\,\sigma^z)|\theta,\phi\rangle = L|\theta,\phi\rangle \tag{22}$$

normalized so that $\langle\theta,\phi|\theta,\phi\rangle = 1$. These two definitions are equivalent: if $z = \cot\frac{\theta}{2}e^{i\phi}$ then

$$|\theta,\phi\rangle = (1+z\bar{z})^{-L}|z\rangle. \tag{23}$$

The relationship between z and (θ,ϕ) has a simple interpretation of projection of the unit sphere onto the complex plane from the north -pole.

The path integral for the matrix element $\langle z_f|e^{-Ht}|z_i\rangle$ is constructed by making use of the resolution of the identity, Eq. (20). Dividing up the time interval $[0,t]$ into n equal segments, inserting the resolution of the identity between each segment and taking the limit $n \to \infty$ leads to the path integral [18]

$$\langle z_f|e^{-Ht}|z_i\rangle = \int_{z_i}^{z_f} \mathcal{D}[\bar{z}(t),z(t)]\exp\{S[\bar{z}(t),z(t)]\}, \tag{24}$$

where the integral runs over all paths $z(t)$, $\bar{z}(t)$ such that $z(0) = z_i$ and $\bar{z}(t) = \bar{z}_f$ [note that $z(t)$ and $\bar{z}(t)$ are *not* required to be complex conjugates]. The action is given by

$$S[\bar{z},z] = \frac{L}{2}\{\ln[1+\bar{z}_f z(t)] + \ln[1+\bar{z}(0)z_i]\} + \int_0^t \left(\frac{L}{2}\frac{\dot{\bar{z}}(s)z(s) - \bar{z}(s)\dot{z}(s)}{1+\bar{z}(s)z(s)} - \mathcal{H}[\bar{z}(s),z(s)]\right) ds, \tag{25}$$

and the Hamiltonian function $\mathcal{H}(\bar{z},z)$ is given by

$$\mathcal{H}(\bar{z},z) = \frac{\langle z|H|z\rangle}{\langle z|z\rangle}. \tag{26}$$

In the large L limit one can make a semiclassical approximation to the path integral

$$\langle z_f|e^{-Ht}|z_i\rangle \approx K \exp\{S[\bar{z}_{\mathrm{cl}},z_{\mathrm{cl}}]\}, \tag{27}$$

where the classical trajectories z_{cl} and $\bar{z}_{\mathrm{cl}}$ satisfy Hamilton's equations

$$\begin{aligned} \dot{\bar{z}}_{\mathrm{cl}} &= \frac{(1+\bar{z}_{\mathrm{cl}}z_{\mathrm{cl}})^2}{L}\frac{\partial\mathcal{H}}{\partial z_{\mathrm{cl}}}, \\ \dot{z}_{\mathrm{cl}} &= -\frac{(1+\bar{z}_{\mathrm{cl}}z_{\mathrm{cl}})^2}{L}\frac{\partial\mathcal{H}}{\partial \bar{z}_{\mathrm{cl}}}. \end{aligned} \tag{28}$$

In order to calculate the form of the classical Hamiltonian Eq. (26) the following results are useful:

$$\langle z|z\rangle = (1+z\bar{z})^L, \tag{29}$$

$$\langle z|\sigma^z|z\rangle = -L(1-z\bar{z})(1+z\bar{z})^{L-1}, \tag{30}$$

$$\langle z|\sigma^x|z\rangle = L(z+\bar{z})(1+z\bar{z})^{L-1}, \tag{31}$$

$$\langle z|\sigma^y|z\rangle = iL(z-\bar{z})(1+z\bar{z})^{L-1}. \tag{32}$$

We next apply these results to solve the dynamics of the EAC model for gene duplication.

IV. MODEL OF ESCAPE FROM ADAPTIVE CONFLICT

We consider a mixed population of sequences of length L (original gene) and $2L$ (duplicated gene), corresponding individuals with single or double copies of the gene undergoing adaptive conflict. For the sequences of length L we take the fitness to be a function of distance from two reference sequences, S^1 and S^2, which are assumed to optimize two different gene functions, namely,

$$f_L(S^i) = f_1[d(S^i,S^1)] + f_2[d(S^i,S^2)], \tag{33}$$

where f_1 and f_2 describe the fitness benefit to the individual from the first and second gene functions, respectively. As the fitness is a function of distance from two reference sequences we require two copies of the σ operators such that the first copy, denoted σ_s, only acts on those sites where two sequences S^1 and S^2 have the same spins, and the second copy, denoted σ_o, only acts on those sites where S^1 and S^2 have the opposite spins. If we denote the number of sites at which S^1 and S^2 have the same spins and the opposite spins by L_s and L_o, respectively, and assume that $S^1 = (\uparrow\uparrow\cdots\uparrow)$ then we can write the Hamming distances of S^i to S^1 and S^2 in terms of the σ operators as follows:

$$d(S^i,S^1)\,||S^i\rangle\rangle = \left(L_s + L_o - \sigma_s^z - \sigma_o^z\right)/2\,||S^i\rangle\rangle, \tag{34}$$

$$d(S^i,S^2)\,||S^i\rangle\rangle = \left(L_s + L_o - \sigma_s^z + \sigma_o^z\right)/2\,||S^i\rangle\rangle, \tag{35}$$

where $||S^i\rangle\rangle$ denotes the vector corresponding to the state S^i. For given S^1 and S^2, L_s and L_o are fixed and σ_s (σ_o) acts in the spin-$L_s/2$ (spin-$L_o/2$) subspace, $V_{L_s/2}$ ($V_{L_o/2}$). The state vector of a single gene, $||S^i\rangle\rangle$, can be expressed as an outer product of a state $|S^i_s\rangle_s$ in the $V_{L_s/2}$ subspace and a state $|S^i_o\rangle_o$ in the $V_{L_o/2}$ subspace. In each subspace, we can use spin coherent states $|z_s\rangle_s$ and $|z_o\rangle_o$ as described in the previous section.

For the sequences of length $2L$, we consider each sequence to be composed of two subsequences of length L, (S^i,S^j), and assume that each subsequence determines the effectiveness of one of the two gene functions, i.e.,

$$f_{2L}(S^i,S^j) = f_1[d(S^i,S^1)] + f_2[d(S^j,S^2)]. \tag{36}$$

For this fitness function we require four copies of the σ operators, denoted σ_{s1}, σ_{o1}, σ_{s2}, and σ_{o2}, which act as follows:

$$\sigma_{s1}||S^i,S^j\rangle\rangle = ||\sigma_s S^i,S^j\rangle\rangle, \tag{37}$$

$$\sigma_{o1}||S^i,S^j\rangle\rangle = ||\sigma_o S^i,S^j\rangle\rangle, \tag{38}$$

$$\sigma_{s2}||S^i,S^j\rangle\rangle = ||S^i,\sigma_s S^j\rangle\rangle, \tag{39}$$

$$\sigma_{o2}||S^i,S^j\rangle\rangle = ||S^i,\sigma_o S^j\rangle\rangle, \tag{40}$$

where $||S^i,S^j\rangle\rangle$ denotes the vector corresponding to the state (S^i,S^j). Again for given S^1 and S^2, $L_{s1} = L_{s2}$ and $L_{o1} = L_{o2}$ are fixed and σ_{s1} (σ_{o1}) and σ_{s2} (σ_{o2}) act in the spin-$L_{s1}/2$ (spin-$L_{o1}/2$), $V_{L_{s1}/2}$ ($V_{L_{o1}/2}$), and in the spin-$L_{s2}/2$ (spin-$L_{o2}/2$) subspace, $V_{L_{s2}/2}$ ($V_{L_{o2}/2}$), respectively. The state vector of a duplicated gene, $||S^i,S^j\rangle\rangle$, can be expressed as an outer product of four states: a state $|S^i_{s1}\rangle_{s1}$ in the $V_{L_{s1}/2}$ subspace, a state $|S^i_{o1}\rangle_{o1}$ in the $V_{L_{o1}/2}$ subspace, a state $|S^i_{s2}\rangle_{s2}$ in the $V_{L_{s2}/2}$ subspace, and a state $|S^i_{o2}\rangle_{o2}$ in the $V_{L_{o2}/2}$ subspace. In each subspace, we can use spin coherent states $|z_{s1}\rangle_{s1}$, $|z_{o1}\rangle_{o1}$, $|z_{s2}\rangle_{s2}$, and $|z_{o2}\rangle_{o2}$ as described in the previous section.

We assume that gene duplication, $S^i \rightarrow (S^i,S^i)$, occurs with a rate ν, and also allow an additive fitness cost c for sequences of length $2L$ to model the cost to the individual of sustaining two rather than one gene copies. Figure 1 shows gene duplication and the final states of duplicated genes after the evolution process. To describe these processes in the Hamiltonian, we need to define a duplication operator, D,

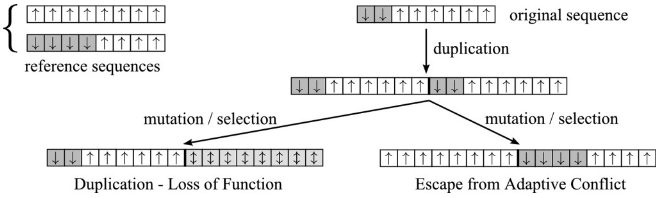

FIG. 1. Process of gene duplication and the final states of duplicated genes.

and two projection operators, I_L and I_{2L}, such that

$$\begin{aligned} DS^i &= (S^i,S^i), \quad D(S^i,S^j) = 0,\\ I_L S^i &= S^i, \quad I_L(S^i,S^j) = 0, \\ I_{2L} S^i &= 0, \quad I_{2L}(S^i,S^j) = (S^i,S^j). \end{aligned} \tag{41}$$

The Hamiltonian then becomes

$$H = H_L + H_{2L} - \nu D, \tag{42}$$

where

$$H_L = -f_L\left(\sigma_s^z,\sigma_o^z\right) + \mu\left(LI_L - \sigma_s^x - \sigma_o^x\right) + \nu I_L, \tag{43}$$

$$\begin{aligned} H_{2L} = &-f_{2L}\left(\sigma_{s1}^z,\sigma_{o1}^z,\sigma_{s2}^z,\sigma_{o2}^z\right)\\ &+\mu\left(2LI_{2L} - \sigma_{s1}^x - \sigma_{o1}^x - \sigma_{s2}^x - \sigma_{o2}^x\right) + cI_{2L}. \end{aligned} \tag{44}$$

For simplicity we will study the case where f_1 and f_2 are linear functions of distance:

$$f_1[d(S^i,S^1)] = A_1[L - 2d(S^i,S^1)], \tag{45}$$

$$f_2[d(S^i,S^2)] = A_2[L - 2d(S^i,S^2)]. \tag{46}$$

In this case $f_L(\sigma_s^z,\sigma_o^z)$ takes the form

$$f_L\left(\sigma_s^z,\sigma_o^z\right) = A_1\left(\sigma_s^z + \sigma_o^z\right) + A_2\left(\sigma_s^z - \sigma_o^z\right), \tag{47}$$

and

$$f_{2L}\left(\sigma_{s1}^z,\sigma_{o1}^z,\sigma_{s2}^z,\sigma_{o2}^z\right) = A_1\left(\sigma_{s1}^z + \sigma_{o1}^z\right) + A_2\left(\sigma_{s2}^z - \sigma_{o2}^z\right). \tag{48}$$

We suppose that initially there are no sequences of length $2L$ in the population and that the population of length L sequences is at equilibrium. Since H_L is a linear combination of the σ operators the equilibrium population is a tensor product of coherent states on the "s" and "o" sectors, $|z_s\rangle_s \otimes |z_o\rangle_o$. Using the methods in Ref. [12] we find

$$z_s = (A_1 + A_2)/\mu + \sqrt{[(A_1 + A_2)/\mu]^2 + 1}, \tag{49}$$

$$z_o = (A_1 - A_2)/\mu + \sqrt{[(A_1 - A_2)/\mu]^2 + 1}, \tag{50}$$

with a mean fitness of the length L sequences:

$$\begin{aligned} \langle f_L\rangle = &\sqrt{(A_1 + A_2)^2 + \mu^2}\, L_s\\ &+ \sqrt{(A_1 - A_2)^2 + \mu^2}\, L_o - \mu L - \nu. \end{aligned} \tag{51}$$

Similarly, we find mean fitnesses of the length $2L$ sequences in the DLoF state and in the EAC state:

$$\begin{aligned} \left\langle f_{2L}^{\mathrm{DLoF}}\right\rangle = &\sqrt{(A_1 + A_2)^2 + \mu^2}\, L_s\\ &+ \sqrt{(A_1 - A_2)^2 + \mu^2}\, L_o - \mu L - c, \end{aligned} \tag{52}$$

$$\left\langle f_{2L}^{\mathrm{EAC}}\right\rangle = \left(\sqrt{A_1^2 + \mu^2} + \sqrt{A_2^2 + \mu^2} - 2\mu\right) L - c. \tag{53}$$

For $c < \nu$, the mean fitness of the length $2L$ sequences is always higher than that of the length L sequences, and therefore as $t \to \infty$ the fraction of length $2L$ sequences in the whole population tends to one. In this regime, two phases of the length $2L$ sequences compete with each other. In case of $A_1 = A_2 = A$, we find the phase boundary in the phase space of $(L_o/L, A/\mu)$:

$$\frac{L_o}{L} = 1 - 2\frac{\sqrt{1 + (A/\mu)^2} - 1}{\sqrt{1 + 4\,(A/\mu)^2} - 1}. \tag{54}$$

This gives the phase diagram shown in the left figure of Fig. 2. For a fixed value of L_o/L, the phase boundary is determined by the condition $\langle f_{2L}^{\mathrm{DLoF}}\rangle = \langle f_{2L}^{\mathrm{EAC}}\rangle$:

$$\begin{aligned} &\sqrt{(A_1 + A_2)^2 + \mu^2}\,\left(1 - \frac{L_o}{L}\right) + \sqrt{(A_1 - A_2)^2 + \mu^2}\,\frac{L_o}{L}\\ &= \sqrt{A_1^2 + \mu^2} + \sqrt{A_2^2 + \mu^2} - \mu, \end{aligned} \tag{55}$$

which can be solved numerically and gives the phase diagram in the right figure of Fig. 2.

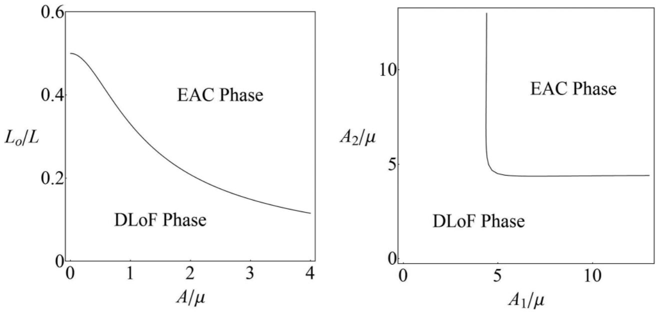

FIG. 2. Parameters $A_1 = A_2 = A$ for the left figure and $L_o/L = 0.1$ for the right figure.

For $c > \nu$, the DLoF phase always has lower fitness than the unduplicated phase so that there is a competition between the unduplicated and EAC phases only. Depending on the parameters $L_o/L, A_1, A_2, \mu, \nu$, and c, the asymptotic state is either the EAC phase or a state dominated by unduplicated sequences of length L. The phase is determined by a competition between the fitness benefits of subfunctionalization and the greater mutational load associated with maintaining two gene copies. The EAC phase is favored when both subfunctions have significant fitness benefits ($A_1, A_2 \gg \mu$) and when the optimal sequences for the two subfunctions are significantly different (large L_o). Since we are interested in the competition between length L and length $2L$ sequences we focus on the dynamics in the $c > \nu$ regime.

As the duplication operator D satisfies $D^2 = 0$ we can expand

$$e^{-Ht} = e^{-(H_L + H_{2L})t} + \nu\int_0^t ds\, e^{-(H_L + H_{2L})s} D e^{-(H_L + H_{2L})(t-s)}. \tag{56}$$

Applying this expansion to the initial state we have

$$\begin{aligned} &e^{-Ht}(|z_s\rangle_s \otimes |z_o\rangle_o)\\ &= e^{\langle f_L\rangle t}(|z_s\rangle_s \otimes |z_o\rangle_o)\\ &\quad + \nu\int_0^t ds\, e^{\langle f_L\rangle(t-s)} e^{-H_{2L}s} D(|z_s\rangle_s \otimes |z_o\rangle_o). \end{aligned} \tag{57}$$

The first and second terms on the right-hand side describe the population of sequences of length L and of length $2L$ at time t, respectively [20,21].

The action of the duplication operator D is problematic. In order to use the coherent states approach we must assume that the population is symmetric, i.e., that any two sequences S^i and S^j which have the same Hamming distance from the reference sequences S^1 and S^2 should have the same population. However, the operator D destroys this symmetry since, for example, duplication of the sequence S^i increases the population of the sequence (S^i,S^i) but does not increase the population of the sequence (S^i,S^j) even though both sequences S^i and S^j have the same Hamming distances from both reference sequences S^1 and S^2.

To deal with this problem we modify the action of the duplication operator to maintain the symmetry. We define the modified duplication operator $\tilde{D}$ by

$$\tilde{D}S^i = \frac{1}{n}\sum_j (S^i,S^j), \tag{58}$$

where the sum runs over all sequences S^j which have the same distances from S^1 and S^2 as those of S^i, and n is the number of such sequences. In terms of representations of $su(2)$, $\tilde{D}$ is the orthogonal projection of D onto the representation of symmetric populations. Since in our model fitness is only a function of distance from the reference sequences the change from D to $\tilde{D}$ does not affect any macroscopic details of the evolution, such as mean fitness.

The action of the modified operator, $\tilde{D}$, on the states $|d)$ is

$$\tilde{D}\,|d) = |d) \otimes |d)\,. \tag{59}$$

The action on a coherent state is therefore found to be

$$\tilde{D}|z\rangle = \tilde{D}\left(\sum_{d=0}^{L} \frac{L!}{(L-d)!d!} z^d\,|d)\right) \tag{60}$$

$$= \sum_{d=0}^{L} \frac{L!}{(L-d)!d!} z^d\,|d) \otimes |d)\,. \tag{61}$$

Rewriting $|d)$ in terms of coherent states and evaluating the sum we find

$$\tilde{D}|z\rangle = \frac{(L+1)^2}{(2\pi)^2} \iint \frac{dz_1 d\bar{z}_1}{(1+z_1\bar{z}_1)^{L+2}} \iint \frac{dz_2 d\bar{z}_2}{(1+z_2\bar{z}_2)^{L+2}} \times (1+z\bar{z}_1\bar{z}_2)^L |z_1\rangle \otimes |z_2\rangle. \tag{62}$$

On the tensor product of coherent states in Eq. (57) the operator $\tilde{D}$ acts as

$$\tilde{D}(|z_s\rangle_s \otimes |z_o\rangle_o) = (\tilde{D}|z_s\rangle_s) \otimes (\tilde{D}|z_o\rangle_o). \tag{63}$$

We will calculate the size of the populations of length L and length $2L$ sequences at time t. From Eq. (57) the population of length L sequences at time t is

$$\begin{aligned} p_L(t) &= ((\cdot|_s \otimes (\cdot|_o) e^{-Ht}(|z_s\rangle_s \otimes |z_o\rangle_o) \\ &= e^{\langle f_L\rangle t}((\cdot|_s \otimes (\cdot|_o)(|z_s\rangle_s \otimes |z_o\rangle_o) \\ &= e^{\langle f_L\rangle t}(1+z_s)^{L_s}(1+z_o)^{L_o}, \end{aligned} \tag{64}$$

and the population of length $2L$ sequences at time t is

$$\begin{aligned} p_{2L}(t) &= \nu \int_0^t ds\, e^{\langle f_L\rangle (t-s)}((\cdot|_{s1} \otimes (\cdot|_{o1} \otimes (\cdot|_{s2} \otimes (\cdot|_{o2}) \\ &\quad \times e^{-H_{2L}s}\tilde{D}(|z_s\rangle_s \otimes |z_o\rangle_o) \\ &= \nu \int_0^t ds\, e^{\langle f_L\rangle (t-s)}((\cdot|_{s1} \otimes (\cdot|_{s2}) e^{-H_{2L}s}(\tilde{D}|z_s\rangle_s) \\ &\quad \times ((\cdot|_{o1} \otimes (\cdot|_{o2}) e^{-H_{2L}s}(\tilde{D}|z_o\rangle_o). \end{aligned} \tag{65}$$

Since the Hamiltonian H_{2L} is linear, its action can be computed separately on each of the four sectors, $s1$, $o1$, $s2$, and $o2$. Defining

$$z_{i\pm} = (A_i/\mu) \pm \sqrt{(A_i/\mu)^2 + 1}, \tag{66}$$

$$\alpha_i(s) = -z_{i-}(1-z_{i-}) + (1+z_{i-})\exp[-\mu(z_{i+}-z_{i-})s], \tag{67}$$

$$\beta_i(s) = (1-z_{i-}) + (1+z_{i-})z_{i-}\exp[-\mu(z_{i+}-z_{i-})s], \tag{68}$$

where $i \in \{1,2\}$, we find

$$\begin{aligned} &((\cdot|_{s1} \otimes (\cdot|_{s2}) e^{-H_{2L}s}(\tilde{D}|z_s\rangle_s) \\ &= \left(\frac{\alpha_1(s)\alpha_2(s) + \beta_1(s)\beta_2(s)z_s}{({z_{1-}}^2+1)({z_{2-}}^2+1)}\right)^{L_s} e^{-L_s cs/L} \\ &\quad \times \exp\left[\frac{1}{2}L_s\mu s(z_{1+}-z_{1-}+z_{2+}-z_{2-}-4)\right], \end{aligned} \tag{69}$$

$$\begin{aligned} &((\cdot|_{o1} \otimes (\cdot|_{o2}) e^{-H_{2L}s}(\tilde{D}|z_o\rangle_o) \\ &= \left(\frac{\alpha_1(s)\beta_2(s) + \beta_1(s)\alpha_2(s)z_o}{({z_{1-}}^2+1)({z_{2-}}^2+1)}\right)^{L_o} e^{-L_o cs/L} \\ &\quad \times \exp\left[\frac{1}{2}L_o\mu s(z_{1+}-z_{1-}+z_{2+}-z_{2-}-4)\right]. \end{aligned} \tag{70}$$

Inserting these results into the integral, Eq. (65), allows us to calculate the population of sequences of length $2L$ at time t.

Evolutionarily, the interesting case is the one in which the double-copy sequences initially have lower fitness than their single-copy parents but through the accumulation of advantageous mutations can raise their mean fitness above that of the single-copy population. The first condition requires that

$$f_L(S^i) - \nu > f_{2L}(S^i,S^i) - c, \tag{71}$$

which reduces to $\nu < c$. The second condition was calculated in Ref. [12] and gives

$$\langle f_L\rangle < \tfrac{1}{2}L\mu(z_{1+}-z_{1-}+z_{2+}-z_{2-}-4) - c. \tag{72}$$

Figure 3 shows the relative size of the populations $p_{2L}(t)$ and $p_L(t)$ for a choice of parameters satisfying these two

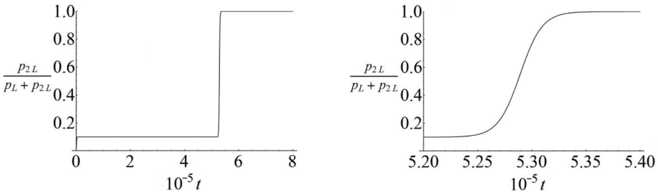

FIG. 3. Fraction of the population in which the gene is duplicated as a function of time. Parameters are $L = 1000$, $L_s = 900$, $A_1 = A_2 = 1.0 \times 10^{-5}$, $\mu = 1.0 \times 10^{-7}$, $\nu = 1.0 \times 10^{-4}$, and $c = 1.0 \times 10^{-3}$.

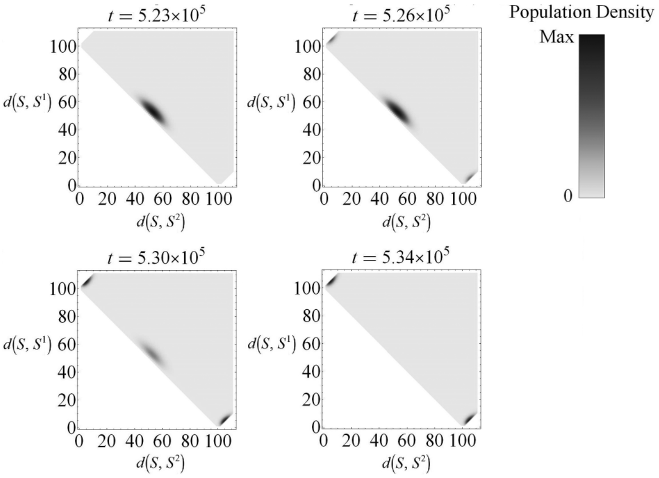

FIG. 4. Distribution of the population as a function of distance from the two reference sequences, $d(S,S^1)$ (vertical axis) and $d(S,S^2)$ (horizontal axis), at various times t. For duplicated sequences the sum of the distributions of each of the two subsequences is shown. Parameters are the same as in Fig. 3. White regions in the diagrams denote points forbidden by the condition $d(S^1,S^2) = L_o$. The distribution changes rapidly at $t^* \approx 5.3 \times 10^5$: for $t = 5.23 \times 10^5$ the distribution is not visibly different from the initial distribution, and by $t = 5.34 \times 10^5$ the distribution is already indistinguishable from its final (asymptotic) form.

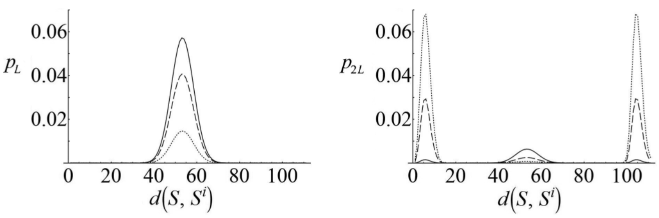

FIG. 5. Population distribution of the length L sequences as a function of distance (left) and population distribution of the two subsequences (two gene copies) of the length $2L$ sequences as a function of distance (right) from each reference sequence ($i = 1,2$). The solid line is at $t = 5.23 \times 10^5$, the dashed line is at $t = 5.28 \times 10^5$, and the dotted line is at $t = 5.30 \times 10^5$.

conditions. Figures 4 and 5 show the evolution of the population of sequences as a function of the distance from the sequences S^1 and S^2 which optimize each gene subfunction. The figures show two distinct regimes: initially the population sustains a small, constant fraction of length $2L$ sequences, until a time t^*, after which the population of length $2L$ sequences starts to exponentially outgrow the population of length L sequences. For $t < t^*$ the length $2L$ sequences have not had time to evolve sufficiently to outperform their length L competitors, and the small, constant fraction of length $2L$ sequences is sustained only by the continuing duplication. For $t > t^*$ a significant number of the length $2L$ sequences have adapted enough to outperform the length L sequences.

This analysis gives us a way to estimate t^*. If the length $2L$ sequences are sustained only by the continuing duplication then the dominant contribution to the integral, Eq. (65), comes from those sequences which have recently been duplicated at time t, i.e., when $s \approx 0$. The integrand in this case is

$$\begin{aligned} &e^{\langle f_L\rangle t}((\cdot|_{s1}\otimes(\cdot|_{o1}\otimes(\cdot|_{s2}\otimes(\cdot|_{o2})\tilde{D}(|z_s\rangle_s\otimes|z_o\rangle_o)\\ &\quad = e^{\langle f_L\rangle t}(1+z_s)^{L_s}(1+z_o)^{L_o}. \end{aligned} \tag{73}$$

The length $2L$ sequences which have had the most time to adapt are those which were duplicated early on, i.e., for which $s \approx t$. If we assume that t is large enough so that $\alpha_i(t)$ and $\beta_i(t)$ are approximately constant then the integrand is

$$\begin{aligned} &((\cdot|_{s1}\otimes(\cdot|_{o1}\otimes(\cdot|_{s2}\otimes(\cdot|_{o2})e^{-H_{2L}t}\,\tilde{D}(|z_s\rangle_s\otimes|z_o\rangle_o)\\ &\quad = \left(\frac{(1-z_{1-})(1-z_{2-})}{(1+z_{1-}{}^2)(1+z_{2-}{}^2)}\right)^L (z_{1-}z_{2-}+z_s)^{L_s}(-z_{1-}-z_{2-}z_o)^{L_o}\exp\left\{\left[\frac{1}{2}L\mu(z_{1+}-z_{1-}+z_{2+}-z_{2-}-4)-c\right]t\right\}. \end{aligned} \tag{74}$$

The crossover between the two regimes occurs approximately when these two contributions are equal, which gives

$$\begin{aligned} t^* &= \left[\frac{1}{2}L\mu(z_{1+}-z_{1-}+z_{2+}-z_{2-}-4)-c-\langle f_L\rangle\right]^{-1}\\ &\quad\times\left[L\ln\left(\frac{(1+z_{1-}{}^2)(1+z_{2-}{}^2)}{(1-z_{1-})(1-z_{2-})}\right)+L_s\ln\left(\frac{1+z_s}{z_{1-}z_{2-}+z_s}\right)+L_o\ln\left(\frac{1+z_o}{-z_{1-}-z_{2-}z_o}\right)\right]. \end{aligned} \tag{75}$$

For the parameters used in Fig. 3 this estimate gives $t^* = 5.3 \times 10^5$, which shows that our approximation is good in this case.

Note that t^* is only positive and finite when the inequality, Eq. (72), is satisfied and tends to infinity when Eq. (72) is satisfied as an equality. As found in Ref. [12], the condition, Eq. (72), as an equality marks a phase transition between two distinct equilibrium phases, one in which sequences of length L dominate the population and one in which sequences of length $2L$ dominate the population. The divergence of t^* at the threshold is thus an example of critical slowing down.

Figure 6 shows t^* plotted as a function of mutation rate and demonstrates that there is an optimum mutation rate at which t^* is a minimum. The existence of an optimum mutation rate is to be expected—too low a mutation rate increases the time necessary for the population of length $2L$ sequences to adapt; too high a mutation rate and the higher mutational load on the length $2L$ sequences destroys their advantage. It is possible to obtain an equation for the optimum mutation rate from $\frac{\partial t^*}{\partial\mu} = 0$, but it is not possible to give a closed form for the optimum mutation rate from this equation.

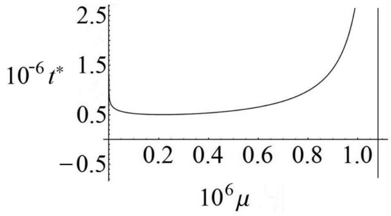

FIG. 6. t^* as a function of mutation rate μ, showing the existence of an optimum mutation rate which minimizes t^*. Other parameters are the same as in Fig. 3. For $\mu > 1.08 \times 10^{-6}$, the inequality, Eq. (72), is no longer satisfied and t^* diverges. The optimum mutation rate occurs when $\mu^* = 2.16 \times 10^{-7}$ per base per generation.

The evolution appears discontinuous around t^*; that is, the peak of the population distribution does not move smoothly but jumps from a point equidistant from S^1 and S^2 (the original pleiotropic gene) to points close to S^1 and S^2 (the subfunctionalized genes). The discontinuity can be understood as follows: over many generations the subfunctionalized genes evolve in a very small subpopulation of duplicated sequences (too small to be visible in the figure), while most of the population retains the original pleiotropic gene. Once the subfunctionalized genes have been discovered this small subpopulation has a fitness advantage and eventually grows to dominate the population. The fact that the crossover at time t^* is determined by the evolution of a small subpopulation implies that in real populations the effects of finite population size are likely to be very significant.

V. DISCUSSION

We have developed and studied a statistical physics model of escape from adaptive conflict for gene duplication, based on the Crow-Kimura quasispecies model. We described the dynamics of a mixed population of individuals with single and double copies of a pleiotropic gene with two functions. The evolution dynamics can be mapped onto the dynamics of a quantum spin chain, which we solved using the spin coherent state path integral.

In the long time limit, there is a competition to dominate in a mixed population between individuals with single genes (the length L sequences) and individuals with duplicated genes (the length $2L$ sequences). In the $c < \nu$ regime, the mean fitness of duplicated genes is always higher than that of individuals with single genes, so that the fraction of individuals with duplicated genes in a mixed population tends to 1 as $t \to \infty$. Furthermore, among individuals with duplicated genes, there is a sharp phase transition between the EAC phase, in which each copy of duplicated genes evolves toward subfunctionalization, and the DLoF phase, in which one copy maintains its pleiotropic function and the other copy undergoes neutral mutation to lose its original function. The phase is determined by a competition between the fitness benefits of subfunctionalization and the greater mutational load associated with maintaining two gene copies. The EAC phase is favored when both subfunctions have significant fitness benefits ($A_1, A_2 \gg \mu$) and when the optimal sequences for the two subfunctions are significantly different (large L_o). That is, subfunctionalization occurs when the cost for gene duplication is smaller than the gene duplication rate, and the fitness benefits of both subfunctions and the distance (L_o) between the sequences optimizing the two functions are larger than some critical values.

In the $c > \nu$ regime, whether individuals with single genes or individuals with duplicated genes dominate in the population depends on the mutation and selection parameters. We chose the mutation and selection parameters such that individuals with duplicated genes initially have lower fitness than their single gene parents but through the accumulation of advantageous mutations can raise their mean fitness higher than those of a single gene population. For these parameters, we showed that there is a sharp change in the composition of the mixed population at time t^*: before t^* the population sustains a small, constant fraction of individuals with duplicated genes; after t^* the population of individuals with duplicated genes starts to exponentially outgrow the population of individuals with single genes and eventually the mixed population consists only of individuals with duplicated genes. We also presented how to estimate t^* and showed that there is an optimal mutation rate at which t^* is a minimum. The crossover at t^* is the result of a small subpopulation of duplicated sequences which develop the two subfunctionalized genes and thereby eventually outperform the rest of the population. A smaller mutation rate increases the time necessary for the subfunctionalized genes to evolve, and a larger mutation rate decreases the fitness benefits of subfunctionalization due to increased mutational load. Thus, the existence of an optimal mutation rate is to be expected.

The values of parameters used in the analysis and figures are based where possible on empirical data. If the fitness value of $A_i L = 0.01$ is taken to represent a 1% fitness benefit to the organism coming from each gene function and each organism produces on average one offspring per generation, then t gives approximately the number of generations. Thus, for the parameter values used in the figures, the model predicts a duplication will take on the order of 10^6 generations to become fixed in the population. Experimental measurement of the duplication rate, ν, is difficult, and estimates can vary by several orders of magnitude according to the methods used [22]. For bacteria and DNA-based viruses, for which the quasispecies model is most directly applicable, typical rates are of the order of 10^{-3}–10^{-5} per gene per generation [23]. Estimates of the spontaneous mutation rates for the same organisms range from 10^{-6}–10^{-10} per base per generation [24]. We have been unable to find any empirical estimates of the fitness cost c of sustaining a duplicated gene, but in our model the possible values of c are constrained by the requirement that the EAC phase be selected. For the choices of parameters above this gives $1.0 \times 10^{-4} < c < 2.0 \times 10^{-3}$.

The results of our quasispecies model apply to infinite populations. However, the behavior of the model and previous theoretical work [25] show that finite size effects such as genetic drift are likely to be very significant in the evolution of subfunctionalization in real populations. Therefore extending the model to finite populations is an important future work.

As far as we know, this is the first presentation of a mathematical model showing the evolutionary dynamics toward the EAC state. In order to focus on the evolutionary dynamics, we considered the simplest fitness landscape, the linear fitness function. Since the spin coherent state path integral can give the dynamics for any symmetric fitness function, the same analysis can also be applied to more complicated cases. We are presently working in the investigation of the effects of epistasis in the evolutionary dynamics toward the EAC state. Duplication is also known to be important in the evolution of transcription factors, where duplication of a transcription factor encoding gene reduces the selective pressure and allows one copy to accumulate more mutations [26,27]. Furthermore, if a regulated gene has several binding sites then there is an evolutionary competition between the maintenance of two weaker binding sites or a single stronger binding site [28]. It should be possible to study such systems by adapting the model presented here.

ACKNOWLEDGMENTS

M.A. was supported by the Catholic University of Korea research fund 2013. J.-M.P. was supported by the Basic Science Research Program through the National Research Foundation of Korea (Grant No. NRF-2013R1A1A2006983).

[1] C. Vogel, M. Bashton, N. D. Kerrison, C. Chothia, and S. A. Teichmann, Curr. Opin. Struct. Biol. **14**, 208 (2004).

[2] R. D. Kouyos, O. K. Silander, and S. Bonhoeffer, Trends Ecol. Evol. **22**, 308 (2007).

[3] M. Garcia-Diaz and T. A. Kunkel, Trends Biochem. Sci. **31**, 206 (2006).

[4] I. G. Choi and S. H. Kim, Proc. Natl. Acad. Sci. USA **103**, 14056 (2006).

[5] M. Lynch, M. O'Hely, B. Walsh, and A. Force, Genetics **159**, 1789 (2001).

[6] J. Zhang, Trends Ecol. Evol. **18**, 292 (2003).

[7] M. Lynch and V. Katju, Trends Genet. **20**, 544 (2004).

[8] S. Ohno, *Evolution by Gene Duplication* (Springer-Verlag, New York, 1970).

[9] J. Piatigorsky and G. Wistow, Science **252**, 1078 (1991).

[10] A. L. Hughes, Proc. R. Soc. B **256**, 119 (1994).

[11] D. L. Des Marais and M. D. Rausher, Nature (London) **454**, 762 (2008).

[12] M. Ancliff and J.-M. Park, J. Stat. Phys. **143**, 636 (2011).

[13] J. F. Crow and M. Kimura, *An Introduction to Population Genetics Theory* (Harper and Row, New York, 1970).

[14] E. Baake, M. Baake, and H. Wagner, Phys. Rev. Lett. **78**, 559 (1997).

[15] D. B. Saakian and C.-K. Hu, Phys. Rev. E **69**, 021913 (2004).

[16] D. B. Saakian and C.-K. Hu, Phys. Rev. E **69**, 046121 (2004).

[17] D. B. Saakian, E. Munoz, C.-K. Hu, and M. W. Deem, Phys. Rev. E **73**, 041913 (2006).

[18] M. Stone, K-S. Park, and A. Grag, J. Math. Phys. **41**, 8025 (2000).

[19] M. Ancliff and J.-M. Park, J. of Korean Phys. Soc. **56**, 891 (2010).

[20] D. B. Saakian, O. Rozanova, and A. Akmetzhanov, Phys. Rev. E **78**, 041908 (2008).

[21] D. B. Saakian, Z. Kirakosyan, and C.-K. Hu, Phys. Rev. E **86**, 031920 (2012).

[22] K. J. Lipinski, J. C. Farslow, K. A. Fitzpatrick, M. Lynch, V. Katju, and U. Bergthorsson, Curr. Biol. **24**, 306 (2011).

[23] R. P. Anderson and J. R. Roth, Ann. Rev. of Microbiol. **31**, 473 (1977).

[24] J. W. Drake, B. Charlesworth, D. Charlesworth, and J. F. Crow, Genetics **148**, 1667 (1998).

[25] M. Lynch and A. Force, Genetics **154**, 459 (2000).

[26] U. Gerland and T. Hwa, J. Mol. Evol. **55**, 386 (2002).

[27] U. Gerland, J. D. Moroz, and T. Hwa, Proc. Natl. Acad. Sci. USA **99**, 12015 (2002).

[28] J. Berg, S. Willmann, and M. Lassig, BMC Evol. Biol. **4**, 42 (2004).