Sharp asymptotics for fixation times in stochastic population dynamics with low mutation probabilities

Alain Cercueil Olivier François

LMC, BP53 38041 Grenoble cedex 9, France
TIMC, Fac. de Médecine, F38706 La Tronche cedex, France

Abstract

This article describes new models in population genetics that extend the neutral Wright-Fisher model by including strong selection and mutation. Fixation times are studied in the limit of small mutation probabilities within the framework of Markov chains with rare transitions. The main result outlines the role of the discrete geometry of the fitness landscape and provides a mean for estimating the expected number of generations for an individual with better fitness value to appear. Some connections to evolutionary algorithms are discussed as well.

keywords: Population genetics. Markov chains with rare transitions. Fixation times.

1 Introduction

Mathematical models in population genetics usually aim at characterizing the gene distribution dynamics in evolving populations quantitatively. Among several goals, these models provide the means to estimate the probabilities of gene fixation, i.e., the condition by which an allele or a group of alleles becomes the only present in a population because of selection.

Both deterministic and stochastic techniques have been introduced for this purpose. Deterministic models rely on the approximation of an infinite population size. In contrast, probabilistic methods deal with finite populations. The methods for computing fixation probabilities or averaged fixation times include the analysis by means of Markov chains using exact computations from generating functions [11], or use approximations by diffusion processes. In this case, the underlying models are known as *Wright-Fisher* models [2, 13, 14]. Although weak selection is sometimes considered, these models are mostly intented to describe neutral evolution.

In contrast, this article focusses on fixation times when selection is a dominant mechanism and when the probability of mutation is small. In this context, Markov chains with rare transitions provide a natural framework for describing evolving populations when the typical pattern includes abrupt appearance of new genotypes and the relative stability of such genotypes in the population (stasis). Such a phenomenon is called evolution by punctuated equilibria, and is related to metastability.

This section presents an introductory model that will be useful as an illustration throughout the article. Several variants of this elementary example could be clearly studied in a similar way. Consider a population of n individuals for which all possible *genotypes* consist of a (large) finite set. In order for gene frequencies to change, either mutation or selection must generally occur. Mutation occurs at a very slow rate and its chance of transmission is small. At the opposite, the selection pressure can be high, and many mutated genes may disappear because their carrier cannot survive. Individuals might be represented as DNA sequences from a molecular region where there appears to be no recombination (e.g., mitochondrial DNA). The molecular region is passed on intact, modulo the effect of mutational substitutions, from parent to offspring. The evolutionary process under consideration can be modelled as follows. An offspring genotype can be created according to

- either the random selection of a parent and a mutation from the parent genotype,

- or the random selection of a parent and the transmission of the parent genotype,

- or the transmission of the "best" genotype in the parental generation (strong selection).

The first event occurs with a small probability p that represents the *mutation probability*. The second event occurs with probability $(1-p)q$ where q is the conditional *transmission probability* given that the mutation does not happen. The last event can be interpreted as follows. If the carrier of a mutation disappears, it is replaced by an offspring whose genotype coincides with the "best" genotype in the previous generation. The word "best" refers to some specific *fitness* or adaptive function. This event happens with probability $(1-p)(1-q)$. The last quantity represents the theoretical fraction of individuals having the same genotype. In our model this fraction would be high.

The article is organized as follows. Section 2 presents the models under interest more specifically, and states the main results about fixation times. Proofs are deferred to Section 3 together with auxillary results. During the recent years, several models of artificial evolution have emerged based on the metaphor of natural evolution [5, 6]. These models are often called *evolutionary algorithms* as they mimic the computational abilities of biological systems. Several connections with these models are discussed, as recent progress has also been made by using techniques based on Markov chains with rare transitions in this framework [3, 4, 7, 8]

2 Main Results

2.1 Models

2.1.1 Definitions

Let E be a finite subset of states (typically a set of genotypes) and X the set of configurations

$$X = E^n, \quad n \geq 1 \tag{1}$$

that can be obtained from the sampling of n genotypes in a population. The elements of X are denoted $x = (x_1, \ldots, x_n)$ where each x_ℓ corresponds to an individual genotype. The integer n corresponds to *the population size.*

Within the set of genotypes, transitions between states can be described by a finite dimensional stochastic matrix

$$\pi = (\pi(a,b))_{a,b \in E}. \tag{2}$$

This matrix contains the transition probabilities corresponding to the substitutions that may occur during the offspring generation.

The *fitness* function is a nonnegative function defined on E. This function is involved in the selection mechanism during the evolution of populations. The *fitness landscape* (f, π) is defined as a graph whose vertices are weighted by the discrete values of the fitness function and the edges by the transition probabilities $\pi(a,b)$. Assuming that two different genotypes cannot be given the same fitness simplifies the description significantly, and we make use of this simplifying hypothesis. Also we denote

$$\hat{x} = \arg\max\{f(x_i),\, i = 1, \ldots, n\}. \tag{3}$$

Selection is a process that tends to maximize the fitness of individuals.

In this article, the evolution of populations is modelled with Markov chains with rare transitions. In such models, transition probabilities are controlled by a small parameter $p > 0$. This parameter represents the rate at which an individual undergoes a mutation. Let $X_t = x$ be the state of the population at generation t. The probability that the population be $X_{t+1} = y$ at time $t+1$ is asymptotically equivalent to

$$p(x,y) \sim c(x,y)\, p^{V(x,y)} \quad \text{as} \quad p \to 0 \tag{4}$$

where $V(x,y) \geq 0$, $c(x,y) \geq 0$ and $\sum_y p(x,y) = 1$ for all $p \geq 0$.

2.1.2 Examples

This section describes two examples and the way by which they can be fitted into the above formalism. The computational properties of the first example have been studied in [8] using large deviations and simulated annealing techniques. The second example corresponds to the evolution mechanisms presented in the introduction.

In the first example, each parent generates a single offspring independent from the others. With probability $1-p$, the offspring genotype is copied from the best parent genotypes (those for which f is maximal). With probability p, the offspring genotype undergoes a mutation from the parent genotype according to the transition matrix π.

More specifically, let x and y be the parent and the offspring population respectively. Define $V(x,y)$ as the *minimal* number of mutations that are necessary to generate population y from population x. The term *minimal* avoids the discussion of coincidences due to mutations that might result in $\hat{x}$. In addition, let $M(x,y)$ be the subset of individuals in population y that correspond to the minimal number of mutations (for such mutations, the offspring genotypes differ from the best genotypes in the previous generation, and the following relationship holds $V(x,y) = \#M(x,y)$). Then, we have

$$\forall x, y \in X \ , \quad p_1(x,y) \sim c_1(x,y) p^{V(x,y)} \quad \text{as} \quad p \to 0, \tag{5}$$

with

$$c_1(x,y) = \prod_{i\colon y_i \in M(x,y)} \pi(x_i, y_i). \tag{6}$$

As a variant, transmission can be added to this model. Conditional to the absence of mutation, either the parent genotype is transmitted to the offspring without modification or the offspring genotype is copied from the best parent genotypes. The first event occurs with probability q and the second event with probability $1-q$. The basic case corresponds to $q=0$. In this situation, we have

$$\forall x, y \in X, \quad p_1(x,y) \sim c_1(x,y) \left(\frac{p}{1-q}\right)^{V(x,y)}, \tag{7}$$

with

$$c_1(x,y) = q^{n_T(x,y)} (1-q)^{n - n_T(x,y) - \hat{n}(x,y)} \prod_{i\colon y_i \in M(x,y)} \pi(x_i, y_i). \tag{8}$$

Here $M(x,y)$ is the subset of individuals in population y that are not in population x, $n_T(x,y)$ is number of individuals in y such that $y_i = x_i$ and $y_i \neq \hat{x}$, and $\hat{n}(x,y)$ is number of individuals in y such that $y_i = x_i$ and $y_i = \hat{x}$.

Describing the second example formally requires more notations. In the second example, each parent may create several offspring as in the Wright-Fisher model. Offspring are created by uniform sampling (with replacement) from the parent population. Again, mutations may occur with probability p. Exact transmission of the genotype is considered as well, and is assumed to happen with probability q conditionally to the absence of mutation. We assume that $q = p^\theta$, $\theta > 0$. Otherwise the offspring genotype is sampled from the best parent genotypes.

Again, let $M(x,y)$ be the subset of individuals in population y that are not in population x, and $V_M(x,y) = \#M(x,y)$ be the number of such individuals. Let $T(x,y)$ be the subset of common members in y and x with genotypes different from $\hat{x}$, and $V_T(x,y) = \#T(x,y)$. Notations M and T stand for mutation and transmission respectively. Finally, let $n(x_i)/n$ denote the frequency of genotype x_i in population x. The evolution can be modelled according to the following Markov chain.

$$\forall x, y \in X, \quad p_2(x,y) \sim c_2(x,y) p^{V_M(x,y) + \theta V_T(x,y)}, \tag{9}$$

with

$$c_2(x,y) = \prod_{i:\, y_i \in T(x,y)} \frac{n(y_i)}{n} \prod_{i:\, y_i \in M(x,y)} \sum_{\ell=1}^{n} \frac{\pi(x_\ell, y_i)}{n}. \tag{10}$$

2.2 Statements of Results

The results presented in this paper bear upon hitting times for the Markovian population dynamics defined in the previous sections. Before giving the result, a set of definitions is needed.

A trajectory γ^E is a sequence of mutations, i.e., a path for the Markov chain of matrix π on E. We denote

$$\pi(\gamma^E) = \pi(a_0, a_1) \ldots \pi(a_{\ell-1}, a_\ell), \quad \ell = \ell(\gamma^E). \tag{11}$$

For all $a, b \in E$, let $d(a,b)$ be the minimal number of transition needed to reach b from a. We denote by Γ^E_{aB} the subset of minimal trajectories from a to $B \subset E$, *i.e.*, the subset of trajectories for which $\ell(\gamma^E) = d(a, B)$.

Theorem 2.1 Consider the stochastic matrix P_1 defined in Section 2.1.2 equation (7) or the stochastic matrix P_2 defined in equation (9). Let (X_t) be the associated Markov chain. Let $a \in E$ and (a) be the uniform population $(a, \ldots, a)$ that consists of n copies of a. Define the subset of X

$$A = \{x \in X,\, f(\hat{x}) > f(a)\}, \tag{12}$$

and the event

$$\mathcal{E} = (X_t \ni a \ \text{ for all } t \leq \tau_{(a)\,A}). \tag{13}$$

For P_1, we have

$$E[\tau_{(a)\,A} \,|\, \mathcal{E}] \sim \frac{(p/1-q)^{-d(a,A)}}{n(1-q)\sum_{\gamma^E \in \Gamma^E_{aB}} \pi(\gamma^E)} \quad \text{as } p \to 0, \tag{14}$$

where $d(a, A)$ is the minimal number of mutation steps needed to reach A from (a), and $B = \{b \in E\,,\, f(b) > f(a)\}$.

For P_2, we have

$$E[\tau_{(a)\,A} \,|\, \mathcal{E}] \sim \frac{p^{-d(a,A)}}{n\sum_{\gamma^E \in \Gamma^E_{aB}} \pi(\gamma^E)} \quad \text{as } p \to 0. \tag{15}$$

Theorem 2.1 is stated conditionally to the realization of the event $\mathcal{E}$. This condition is more a convenient technical assumption than a restrictive hypothesis. Indeed, one has

$$\text{Prob}(\mathcal{E}) \sim 1, \quad \text{as } p \to 0, \tag{16}$$

if the chain is started from (a).

The computation of $d(a, A)$ in equation (14) can be achieved from the knowledge of the fitness landscape only, i.e., the values of the fitness function for all vertices in the mutation graph. Then, $d(a, A)$ represents the distance from the genotype a to the subset of genotypes of better adaptation, and has a natural "geometric" interpretation.

The above result has an obvious interpretation in term of fixation times for the evolutionary dynamics defined from P_1 or P_2. In the models presented in section 2.1.2, populations most often consist of copies of a single genotype. Mutation and transmission are mechanisms that enable new genotypes to appear and survive but for few generations only. When a better adapted genotype appears, it becomes dominant in the population abruptly and fixation occurs. For small mutation probabilities, equation (14) is a good estimation of the average fixation time.

3 Proofs

This section is devoted to the proofs of our main results. A general result for Markov chains with rare transitions is presented first. Then, specific results regarding the examples of section 2.1.2 are stated. An additional result will be stated in Section 3.3.

We start with some definitions borrowed from [9]. The objects γ, Γ_{xy}, $c(\gamma)$ will have definitions relative to π similar to those of γ^E, Γ^E_{ab}, $\pi(\gamma^E)$. While the first set of definitions corresponds to the individual level, this new set corresponds to the population level.

Consider a trajectory γ of a Markov chain of matrix P satisfying equation (4)

$$\gamma : \, x_0 \to x_1 \to \ldots \to x_\ell ,$$

where $\ell = \ell(\gamma)$ is the length of γ. We set

$$p(\gamma) = p(x_0, x_1) \ldots p(x_{\ell-1}, x_\ell) \sim c(\gamma) p^{V(\gamma)}. \tag{17}$$

For all $x, y \in X$, let

$$W(x,y) = \min\{V(\gamma) \, ; \, \gamma : \, x_0 = x \to x_1 \to \ldots \to x_\ell = y \, , \, \ell \geq 1\}, \tag{18}$$

and, for $A \subset X$,

$$W(x,A) = \min\{W(x,y) \, ; \, y \in A\}. \tag{19}$$

We denote by Γ_{xA} the subset of minimal trajectories from x to A, *i.e.*, the subset of trajectories for which $V(\gamma) = W(x,A)$.

Let $x \in X$ and $A \subset X$. The hitting time of A starting from x is

$$\tau_{xA} = \min\{t \geq 0 \, ; \, X_t \in A \, , \, X_0 = x\}. \tag{20}$$

Theorem 3.1 Consider a stochastic matrix P satisfying equation (4). Assume that

1. there exists a unique x^* in X such that $p(x^*, x^*) \sim 1$ as p goes to 0;

2. there is no closed trajectory γ such that $V(\gamma) = 0$ and $\ell(\gamma) \geq 2$.

Then, for all $A \not\ni x^*$, we have

$$E[\tau_{x^*A}] \sim \left(\sum_{\gamma \in \Gamma_{x^*A}} c(\gamma) \prod_{x \in \gamma \setminus x^* \cup A} \frac{1}{1 - p(x,x)} \right)^{-1} p^{-W(x^*,A)} \quad \text{as } p \to 0. \tag{21}$$

In addition, we have

$$Var[\tau_{x^*A}] \sim E[\tau_{x^*A}]^2 \quad \text{as } p \to 0. \tag{22}$$

Under the theorem's assumption, the front term in equation (21) converges to a constant as p goes to zero. The first result stated in Theorem 3.1 establishes that the average hitting of time of an arbitrary subset A becomes proportional to $p^{-W(x^*,A)}$, and the proportionality coefficient is known.

The constant and the order parameter $W(x^*, A)$ can hardly be made explicit in general situations. Nevertheless, this result will be applied to studying hitting times for the stochastic population dynamics defined by P_1 and P_2.

3.1 Proofs of the main results

Let P denote the transition matrix associated with the Markov chain (X_t) defined in equation (4). The Markov chain (X_t) is defined on a finite state space X, and the elements of X can be labelled $1, \ldots, m$. Up to this point, no distinction will be made between X and the set of labels $\{1, 2, \ldots, m\}$.

Let A be a subset of X. Let T denote the vector $(\tau_{yA})_{y \notin A}$ and P_A the matrix whose components are the $p(x,y)$ for $x, y \notin A$. According to a classical result [12], T is given by

$$T = Q_A^{-1}(-1), \tag{23}$$

where $Q_A = P_A - I$ and (-1) denotes the vector with all components equal to -1.

Let Δ^A be the determinant of Q_A

$$\Delta^A = \det(Q_A) \tag{24}$$

and let Δ^A_{xy} denote the minor (x,y) of Q_A. With these notations, we have

$$E[\tau_{xA}] = -\sum_{y \notin A} (-1)^{x+y} \frac{\Delta^A_{xy}}{\Delta^A}. \tag{25}$$

Let x and y be two elements of X such that $x \notin A$. $\tilde{\Gamma}^A_{xy}$ will denote the set of paths

$$\gamma : x \to x_1 \to \ldots \to x_k \to y \,, \quad x_i \notin A. \tag{26}$$

Then $W^A(x,y)$ is defined as

$$W^A(x,y) = \min\{V(\gamma); \gamma \in \tilde{\Gamma}^A_{xy}\} \tag{27}$$

and $\Gamma^A_{xy} = \{\gamma \in \tilde{\Gamma}^A_{xy}\,;\, V(\gamma) = W^A(x,y)\}$.

Note that for any subset $Y \subset X$ and Q any square matrix on Y, we have

$$\det(Q) = \sum_{\sigma \in S(Y)} \epsilon(\sigma) \prod_{y \in Y} q(y, \sigma(y)) \tag{28}$$

where $S(Y)$ is the symmetric group of Y and $\epsilon(\sigma)$ is the signature of σ.

For y_0 in Y and $\sigma \in S(Y)$, consider the orbit of y_0

$$\gamma = y_0 \to \sigma(y_0) \to \ldots \to \sigma^j(y_0) = y_0 \tag{29}$$

then, we have

$$\det(Q) = \sum_{\gamma \in \tilde{\Gamma}_{y_0 y_0}} (-1)^{\ell(\gamma)+1} q(\gamma) \sum_{\sigma \in S(Y \setminus \gamma)} \epsilon(\sigma) \prod_{y \in Y \setminus \gamma} q(y, \sigma(y)). \tag{30}$$

3.2 Proof of Theorem 3.1.

The proof is decomposed into several steps.

Step 1. Computation of Δ^A - case 1: $x^* \in A$.

We have

$$\Delta^A = \prod_{y \notin A} (p(y,y) - 1) + o(1) \tag{31}$$

Proof. Take σ equal to identity in equation (28), then the corresponding term in the sum is $\prod_{y \notin A} p(y,y) - 1$. Under hypothesis (H1), we have

$$p(y,y) = 1 + o(1) \quad \text{iff} \quad y = x^*. \tag{32}$$

Hence the term $\prod_{y \notin A} p(y,y) - 1$ is of order 0.

If σ differs from the identity, there exists an $x_0 \notin A$ such that $\sigma(x_0) \neq x_0$. Then let γ be the orbit of x_0. Since $\ell(\gamma) \geq 2$, hypothesis (H2) implies that $V(\gamma) > 0$. Hence the corresponding term in equation (30) is of order higher (or equal) than 1. $\diamond$

Step 2. Computation of Δ^A - case 2: $x^* \notin A$.

We have

$$\Delta^A = \sum_{\gamma \in \tilde{\Gamma}_{x^* A}} (-1)^{\ell(\gamma)} p(\gamma) \Delta^{A \cup \gamma} \tag{33}$$

Proof. let $(C_x)_{x \notin A}$ denote the columns of Q_A. Let C_A denote a vector whose elements are equal to $\sum_{y \in A} p(x,y)$ with $x \notin A$. The operation that replaces the column C_{x^*} by $\sum_{y \notin A} C_y$ leaves the determinant unchanged. let Q'_A be the matrix obtained from Q_A according to this transformation. Since the matrix P is stochastic, we have

$$\sum_{y \notin A} C_y = -C_A \tag{34}$$

and

$$q'_A(x^* \to \ldots \to x_\ell \to x^*) = -q_A(x^* \to \ldots \to x_\ell \to A) \tag{35}$$

Replacing this equality in equation (30) completes the proof of the result. $\diamond$

Step 3. Computation of Δ^A_{xy}.

We have

$$(-1)^{x+y}\Delta^A_{xy} = \sum_{\gamma\in\tilde{\Gamma}^A_{yx}} (-1)^{\ell(\gamma)} p(\gamma)\Delta^{A\cup\gamma}. \tag{36}$$

Proof. Let Q''_A be the matrix obtained from Q_A by setting the all coefficients in the line x and all coefficients in the column y equal to 0, except for the coefficient corresponding to (x, y) which is set up to 1. Obviously, we have

$$\det(Q''_A) = (-1)^{x+y}\Delta^A_{xy}. \tag{37}$$

Let $\gamma' = x \to x_1 \to \ldots \to x_\ell \to x$ be a path in $\tilde{\gamma}^A_{yy}$. Since $q''_A(x, t) = 0$ for $t \neq x$, we have $q''_A(\gamma') \neq 0$ if and only if $x_1 = y$. Then,

$$q''_A(\gamma') = q_A(y, x_2)\ldots q_A(x_\ell, x) = q_A(y \to x_2 \to \ldots \to x_\ell \to x). \tag{38}$$

Reporting this in equation (30) leads to the result. $\diamond$

Step 3. Computation of mean fixation times.

Now, we are ready for the final step in proving Theorem 3.1. First, notice that

$$\Delta^A = \sum_{\gamma\in\tilde{\Gamma}_{x^*A}} (-1)^{\ell(\gamma)} p(\gamma)\,\Delta^{A\cup\gamma} \tag{39}$$

$$= \sum_{\gamma\in\tilde{\Gamma}_{x^*A}} (-1)^{\ell(\gamma)} c(\gamma) p^{V(\gamma)}\,\Delta^{A\cup\gamma} \tag{40}$$

where, according to equation 31,

$$\Delta^{A\cup\gamma} = \prod_{y\notin A\cup\gamma} (p(y, y) - 1) + o(1). \tag{41}$$

Therefore, we have

$$\Delta^A = \sum_{\gamma \in \tilde{\Gamma}_{x^*A}} (-1)^{\ell(\gamma)} c(\gamma) p^{V(\gamma)} \left(\prod_{y \notin A \cup \gamma} (p(y,y) - 1) + o(1) \right). \quad (42)$$

The dominant term in the above sum is obtained by summing over all terms for which $V(\gamma)$ is minimal. This yields $V(\gamma) = W(x^*, A)$ and the sum runs over all $\gamma \in \Gamma_{x^*A}$. Replacing in equation (42), we have

$$\Delta^A = \sum_{\gamma \in \Gamma_{x^*A}} (-1)^{\ell(\gamma)} c(\gamma) p^{W(x^*,A)} \prod_{y \notin A \cup \gamma} (p(y,y) - 1) (1 + o(1)). \quad (43)$$

Since the product $\prod_{y \notin A \cup \gamma} (p(y,y) - 1)$ contains $m - |A| - \ell(\gamma)$ terms (the end of γ is in A), we have

$$\prod_{y \notin A \cup \gamma} (p(y,y) - 1) = (-1)^{m-|A|-\ell(\gamma)} \prod_{y \notin A \cup \gamma} (1 - p(y,y)), \quad (44)$$

and

$$\Delta^A = p^{W(x^*,A)} (-1)^{m-|A|-1} \sum_{\gamma \in \Gamma_{x^*A}} c(\gamma) \prod_{y \notin A \cup \gamma} (1 - p(y,y)) (1 + o(1)). \quad (45)$$

Proceeding with all other minors in a similar way leads to the following result

$$(-1)^{m-|A|-1} \Delta^A_{x^*x^*} = \prod_{y \notin A \cup x^*} (1 - p(y,y)) + o(1). \quad (46)$$

If $x \neq x^*$, Δ^A_{xx} is given by

$$\Delta^A_{xx} = \Delta^{A \cup x}. \quad (47)$$

If $x = x^*$ or $y = x^*$, x^* is in $A \cup \gamma$ for all γ in Γ^A_{yx}. Then, we have

$$|\Delta^A_{xy}| = \left(\sum_{\gamma \in \Gamma^A_{yx}} c(\gamma) \prod_{z \notin A \cup \gamma} (1 - p(z,z)) \right) p^{W^A(y,x)} (1 + o(1)). \quad (48)$$

and

$$(-1)^{m-|A|-1}(-1)^{+x+y}|\Delta^A_{xy}| = \Delta^A_{xy}. \tag{49}$$

In order to compute the mean hitting time, recall that

$$E[\tau_{x^*A}] = -\sum_{y\notin A}(-1)^{x^*+y}\frac{\Delta^A_{yx^*}}{\Delta^A}. \tag{50}$$

Since $p(x^*, x^*) = 1 + o(p)$, we have $V(x^*, y) > 0$ for all $y \neq x^*$ such that $c(x^*, y) \neq 0$, and

$$W^A(x^*, y) \geq W(x^*, y) > 0. \tag{51}$$

Hence, we obtain

$$\sum_{y\notin A}(-1)^{x^*+y}\Delta^A_{yx^*} = \Delta^A_{x^*x^*} + o(1). \tag{52}$$

An interesting consequence is

$$E[\tau_{x^*A}] = -\frac{\Delta^A_{x^*x^*}}{\Delta^A}(1 + o(1)). \tag{53}$$

Finally, we have

$$\begin{aligned} E[\tau_{x^*A}] &= \frac{\prod_{y\notin A\cup x^*}(1-p(y,y))}{p^{W(x^*,A)}\sum_{\gamma\in\Gamma_{x^*A}}\prod_{y\notin A\cup\gamma}(1-p(y,y))c(\gamma)}(1+o(1)) && (54)\\ &= Kp^{-W(x^*,A)}(1+o(1)) && (55) \end{aligned}$$

where

$$K = \left(\sum_{\gamma\in\Gamma_{x^*A}}\frac{c(\gamma)}{\prod_{y\in\gamma\setminus(x^*\cup A)}1-p(y,y)}\right)^{-1}. \tag{56}$$

This concludes the first part of Theorem 3.1. ◇

Step 4. Computation of variances.

The second part of Theorem 3.1 is devoted to the computation of variances. Let V be the vector of components equal to $E[\tau^2_{xA}]$, $x \notin A$. Some basic linear algebra shows that

$$V = 1 + P_AV + 2P_AT = P_AV + 2T - 1 \tag{57}$$

and

$$V = (Q_A^{-1} + 2Q_A^{-2})1 \tag{58}$$

The coefficient (x, y) in Q_A^{-2} is equal to

$$Q_A^{-2}(x, y) = (-1)^{x+y} \sum_{z \notin A} \Delta_{zx}^A \Delta_{yz}^A (\Delta^A)^{-2}. \tag{59}$$

The term $\Delta_{zx}^A \Delta_{yz}^A$ is of order $O(p^{W^A(x,z)+W^A(z,y)})$ for $z \neq x$ and $z \neq y$. For all $z \neq x^*$, we have $\Delta_{zx^*}^A = o(1)$, and if $z = x^*$, we have $\Delta_{yx^*}^A = o(1)$ for all $y \neq x^*$. Finally, for $x = x^*$, the only term $\Delta_{zx}^A \Delta_{yz}^A$ of order 0 is $\Delta_{x^*x^*}^A \Delta_{x^*x^*}^A$, and hence

$$E[\tau_{x^*A}^2] = 2\left(\frac{\Delta_{x^*x^*}^A}{\Delta^A}\right)^2 (1 + o(1)) = 2E[\tau_{x^*A}]^2 (1 + o(1)) \tag{60}$$

and

$$Var(\tau_{x^*A}) = E[\tau_{x^*A}]^2 (1 + o(1)). \tag{61}$$

3.3 Proof of Theorem 2.1

The proof of Theorem 2.1 is by far shorter than the previous one. The result arises from Theorem 3.1 directly. The trajectories that are the most probable with respect to P_1 or P_2 have been described in details in [8].

Let us start with P_1. In particular, the most probable trajectory from a uniform population (a) to a population containing an individual "better" than a consists in keeping $n-1$ individuals equal to a and letting one individual evolve from a to b with adaptive value "better" than a. For each path, γ^E de E, there are n paths γ in X that correspond to the n choices for the offspring of mutation (say i). Along a path

$$\gamma : (a) \to \ldots \to x^k \to x^{k+1} \to \ldots \to y \in A,$$

the transition probabilities satisfy

$$p_1(x^k, x^{k+1}) = \pi(x_i^k, x_i^{k+1})p + o(p). \tag{62}$$

If $k \neq 0$, we have $x_i^k \neq a$ and $x_j^k = a$ for $j \neq i$. Then,

$$p_1(x^k, x^k)) = q + o(1). \tag{63}$$

Conditionally to $\mathcal{E}$, we apply Theorem 3.1 with $x^* = (a)$

$$E[\tau_{(a)A}|\mathcal{E}] \sim \frac{(p/1-q)^{-d(a,B)}}{n(1-q)\sum_{\gamma^E \in \Gamma^E} \pi(\gamma^E)}. \tag{64}$$

A similar result holds for P_2. Again, the most probable trajectory from a uniform population (a) to a population containing an individual "better" than a consists in keeping $n-1$ individuals equal to a and letting one individual go from a to b with adaptive value "better" than a. For each path, γ^E de E, there are $n^{\ell(\gamma^E)}$ paths γ in X corresponding to the n choices for the offspring of mutation at each of the $\ell(\gamma^E)$ steps.

Along he path

$$\gamma : (a) \to \ldots \to x^k \to x^{k+1} \to \ldots \to y \in A,$$

let $x^k_{i_k}$ denote the offspring of mutation at step k. The transition probability satisfies

$$p_2(x^k, x^{k+1}) = \pi(x^k_{i_k}, x^{k+1}_{i_{k+1}})p/n + o(p) \tag{65}$$

if $k \neq 0$ (because in this case $n(x^k_{i_k}) = 1$) and

$$p_2((a), x^1) = \pi(a, x^1_{i_1})p + o(p) \tag{66}$$

if $k = 0$ (in this case we have $n(x^k_{i_k}) = n(a) = n$).

If $k \neq 0$, we have $x^k_{i_k} \neq a$ and $x^k_j = a$ for $j \neq i_k$. Then,

$$p_2(x^k, x^k)) = 1 + o(1). \tag{67}$$

Conditionally to $\mathcal{E}$, we apply Theorem 3.1 with $x^* = (a)$ and obtain

$$E[\tau_{(a)A}|\mathcal{E}] \sim \frac{p^{-d(a,B)}}{n\sum_{\gamma^E \in \Gamma^E} \pi(\gamma^E)}. \tag{68}$$

3.4 An additional result

This section presents an additional result that describes the probability of hitting B before A. This result emphasizes the exponential-like behaviour of hitting times for small mutation probabilities.

Let A and D be two subsets of X and x an element of X such that $x \notin D$. The hitting time of A starting from x before entering D is

$$\tau_{xA}^{D} = \min\{t \geq 0; X_0 = x; X_t \in A; x_1, \ldots, X_{t-1} \notin A \cup D\}. \tag{69}$$

The expectation of $\tau_{x^*A}^{D}$ is given by

$$E[\tau_{x^*A}^{D}] = \left(p^{W^D(x^*,A)} \sum_{\gamma \in \Gamma_{x^*A}^{D}} \frac{c(\gamma)}{\prod_{y \in \gamma \setminus (x^* \cup A)} (1 - p(y,y))} \right)^{-1} (1 + o(1)). \tag{70}$$

Theorem 3.2 Let A and B be two non intersecting subsets in X such that $x^* \notin A$ and $x^* \notin B$. We have

$$\mathrm{Prob}(\tau_{x^*A} < \tau_{x^*B}) = \frac{E[\tau_{x^*B}^{A \cup B}]}{E[\tau_{x^*A}^{A \cup B}] + E[\tau_{x^*B}^{A \cup B}]} (1 + o(1)). \tag{71}$$

Proof. let R_A (resp. R_B) be the vector whose components are $p(y, A)$ (resp $p(y, B)$) for $y \notin A \cup B$. The Markov chain (X_t) is modified so that if $X_t \in A \cup B$ then $X_{t+1} = X_t$. The transition matrix of the modified chain is equal to

$$\tilde{P} = \begin{pmatrix} P_C & R_A & R_B \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{72}$$

By a standard argument, we have

$$\mathrm{Prob}(\tau_{xA} < \tau_{xB}) = \lim_{n \to \infty} \tilde{p}^n(x, A), \tag{73}$$

and

$$\mathrm{Prob}(\tau_{.A} < \tau_{.B}) = -(P_{A \cup B} - I)^{-1} R_A. \tag{74}$$

Replacing $(P_{A \cup B} - I)^{-1}$ yields

$$\begin{aligned} \mathrm{Prob}(\tau_{x^*A} < \tau_{x^*B}) &= -\frac{\sum_{y \notin A \cup B} (-1)^{x^*+y} \Delta_{yx^*}^{A \cup B} p(y, A)}{\Delta^{A \cup B}} && (75) \\ &= -\frac{\sum_{y \notin A \cup B} \sum_{\gamma \in \tilde{\Gamma}_{x^*y}^{A \cup B}} (-1)^{\ell(\gamma)} p(\gamma) p(y, A) \Delta^{A \cup B \cup \gamma}}{\Delta^{A \cup B}} && (76) \end{aligned}$$

Now, transform $\gamma = x^* \to \ldots \to y$ into $\gamma' = x^* \to \ldots \to y \to z \in A$. The new path satisfies

$$p(\gamma') = p(\gamma)p(y, z). \tag{77}$$

Therefore, we have

$$\mathrm{Prob}(\tau_{x^*A} < \tau_{x^*B}) = \frac{\sum_{\gamma' \in \tilde{\Gamma}^{A \cup B}_{x^*A}} (-1)^{\ell(\gamma')} p(\gamma') \Delta^{A \cup B \cup \gamma'}}{\sum_{\gamma' \in \tilde{\Gamma}^{A \cup B}_{x^*A \cup B}} (-1)^{\ell(\gamma')} p(\gamma') \Delta^{A \cup B \cup \gamma'}}. \tag{78}$$

The result follows by splitting the denominator into two inverse expectations. $\diamond$

This result provides an explanation why natural or simulated population processes may not follow their most probable trajectories (that lead to the closest individual with better adaptive value). Instead, shortcuts are always possible. Regarding simulated evolution procedures in optimization, this result also suggests that maintaining short independent parallel runs of population algorithms might be more efficient than keeping a single long run given the same computational resource.

3.5 Connections to Evolutionary Algorithms

During the recent years, several models of evolutionary algorithms have been studied within the simulated annealing framework[3, 4, 7, 8]. Simulating these Markovian models leads the user to the observation of metastable states, and long stasis during which few improvements of the solutions can be obtained. Section 3.5.1 presents an overview of recent results for these particular Markov chains with rare transitions. These results use the formalism of large deviations. Most of them actually required trajectorial techniques of proof that differ from the algebraic techniques involved in section 3. Another difference arises from the very nature of the results. Computational properties of evolutionary algorithms are usually investigated with mutation probabilities depending on a positive temperature parameter that slowly decreases to zero. While the above references describe simulated annealing-like theories, the next section extracts and restates the results that bear on hitting time of an optimal population (i.e., a population containing an individual genotype of optimal fitness).

In contrast to the statements presented in section 2.2, the results obtained via the large deviations formalism are rough logarithmic equivalents. The implications of precise equivalents in implementing evolutionary algorithms will be discussed in Section 3.5.2.

3.5.1 A Brief Overview of Evolutionary Algorithms Results

In describing the dynamics of an ergodic Markov chain with rare transitions, subsets called *cycles* play a central role [9], [10], [3]. A subset $C \subset X$ is a cycle if either it consists of a single population, or for all x, y in C, the expected number of "cyclic" visits to x followed by y before exiting from C is exponential

$$E[N_{xy}(C)] \approx p^{-K_{xy}(C)} \ , \quad K_{xy}(C) > 0 \tag{79}$$

(the symbol $\approx$ means that the relationship is a logarithmic equivalent). As a consequence, a cycle should be explored systematically before the chain exits and proceeds with an other cycle. Here is a trajectorial definition of cycles, which is more amenable to a mathematical analysis [15], [3]. For all $x, y \in X$, $x \neq y$, and each trajectory γ, define the *elevation* as

$$H(\gamma) = \max_{0 \leq k < r} \{V(x_k) + V(x_k, x_{k+1})\}, \tag{80}$$

with the maximum taken over all vertices in γ, and

$$V(x) = \lim_{p \to 0} \log \mu_p(x) / \log p, \tag{81}$$

where μ_p is the (unique) invariant probability distribution of the chain. Let $H(x, y)$ be the lowest possible value of $H(\gamma)$ over all self-avoiding trajectories γ from x to y. The quantity $H(x, y)$ is called the *communication altitude*. Now, let $\lambda \geq 0$ and $V_\lambda = \{x \in X \ ; \ V(x) \leq \lambda\}$. Say that x and y *communicate* at height λ in V_λ if $H(x, y) \leq \lambda$. A subset $C \subset X$ is a cycle if all populations are able to communicate at height λ for some $\lambda > 0$.

While the hierarchy of cycles may be extremely complex in general, a remarkable fact is that, for large population sizes, a cycle C which does not contain the optimal population (a^*) reduces to a single population. This result actually holds for the induced chain on the set of uniform populations. As evolutionary algorithms can reach uniform populations at null cost $(V(x,(x_*)) = 0)$, the induced chain nevertheless gives a right picture of the dynamics for small mutation probabilities. For genetic algorithms, the critical population size n_* has been estimated by Cerf [3]. For our basic process, the critical value is lower than

$$n_* = \max_{a \neq a^*} d(a, a^*). \tag{82}$$

Now, denote by τ_C the exit time of the subset C (the hitting time of $\bar{C}$). The expected value can be computed as

$$E[\tau_C] \approx p^{-H_e(C)}, \tag{83}$$

where, according to [15], $H_e(C)$ is *the exit height of* C defined as

$$H_e(C) = \max_{x \notin C} \min_{y \in C} \{H(x,y) - V(x)\}. \tag{84}$$

Let τ_* denote the hitting time of the absolute optimum. The expected value of τ_* can therefore be approximated as

$$E[\tau_*] \approx p^{-H_1}, \tag{85}$$

where

$$H_1 = \max\{H_e(C) \ ; \ C \text{ cycle not intersecting } \mathcal{V}_*\}, \tag{86}$$

and $\mathcal{V}_* = \arg\min V$. For the both algorithms, H_1 can be given as

$$H_1 = \max_{x \neq (a^*)} \left\{ H(x,(a^*)) - V(x) \right\}, \tag{87}$$

whenever $n > n_*$. As far as our basic example is concerned, [8] shows that

$$H_1 = \max_{a \neq a^*} \min_{b : f(b) > f(a)} d(a,b). \tag{88}$$

In other words, H_1 is the minimal number of mutations required for a genotype to exit from any local (non global) minimum in the fitness landscape. This quantity plays the same role in the implementation of an algorithm as the critical depth in simulated annealing procedures [10].

3.5.2 Algorithmic Implications of the Results

This section is devoted to the application of Theorem 2.1 to the optimization algorithm that corresponds to our basic model P_1 ($q = 0$). An elistist version of the algorithm can be implemented easily. In such a case, we have

$$\text{Prob}(\mathcal{E}) = 1. \tag{89}$$

As often assumed in genetic algorithms, the set of genotypes E can be taken as the set of bit strings of length k, i.e.,

$$E = \{0, 1\}^k. \tag{90}$$

Assume that mutations occur randomly, i.e., every genotype can be reached by mutation in a single step. The new genotype is randomly chosen among the 2^k possible genotypes in E, and we say that the fitness landscape is fully connected. For any pair (a, b) in E^2, we have $\pi(a, b) = 2^{-k}$. Then Theorem 3.1 can be applied to the Markov chain defined by equation (6) that models the basic mutation-selection evolutionary algorithm studied in [8]. Taking $q = 0$, mean hitting times are given by

$$E[\tau_{(a)A}] \sim \frac{2^k}{nm_a} p^{-1} \tag{91}$$

where

$$A = \{x \in X \, , \, f(\hat{x}) > f(a)\}, \tag{92}$$

and m_a is the number of genotypes in E with adaptive values lower than $f(a)$

$$m_a = \#\{b \in E, \, f(b) > f(a)\} = \#B. \tag{93}$$

Equation (85) indicated that fully connected structures become good in the asymptotical settting (because $H_1 = 1$). In contrast, equation (91) shows that this may be true for very small mutation probabilities only. Indeed hitting times become proportional to the size of E and the method has the same order of performances as enumeration. Fully connected structures are usually precluded in implementing a optimization algorithm, and some kind of local search is always considered within the evolutionary procedure.

For instance, single bit mutation is a widely used example of a mutation operator [5]. When mutation occurs, a bit is randomly chosen among the k possible and flipped. In this case, the minimal number of mutations required to change the state a into the state b is the number of bits of b that differ from those of a. This quantity is also known as the *Hamming distance* between a and b.

If $d(a,b) = d$, d steps are required to go from a to b. The d associated mutations may happen in any of the $d!$ orders. For each step, we have $\pi(a,b) = 1/k$. Finally, mean hitting times are given by

$$E[\tau_{(a)A}] \sim \frac{(p/k)^{-d(a,A)}}{n\, m_a\, d(a,A)!} \tag{94}$$

where

$$m_a = \#\{b \in E,\ f(b) > f(a)\, d(a,b) = d(a,A)\}. \tag{95}$$

Again, this result is far more accurate than equation (85). In addition, the condition that the population size be greater than a threshold value (which is necessary and sufficient in the simulated-like framework) is not critical in studying average fixation times. This explains why the algorithm may work well even when population sizes are small.

4 Conclusion

This article has presented new models of evolving populations that can be viewed as Markov chains with rare transitions. In such models, the probability of a transition from a parent to the offspring is controlled by a small disorder parameter. In these models, the parameter is the probability that a genotype undergoes a mutation. In living organisms, these probabilities are usually measured in the range $10^{-4} - 10^{-8}$.

The models take their inspiration from simulated evolution where the goal is optimizing an objective function. A basic (and efficient) procedure based on a fraction of elitism has been modified so that it includes exact transmission of genotypes as well as random sampling. The modified model can actually be considered as a natural extension of the classical Wright-Fisher model ($p = 0, q = 1$).

Our main result has described the hitting times of populations of better adaptive values and hence fixation times (or punctuated equilibria). There is a close relationship between our results and those obtained in the large deviations/simulated annealing framework. Both approaches outline the role of the discrete geometry of the fitness landscape. However there are important differences as well. Our results are based on algebraic instead of trajectorial techniques. As a consequence, we were able to establish sharp asymptotics instead of rough logarithmic equivalents.

References

[1] O. Catoni. Simulated annealing algorithms and Markov chains with rare transitions. Lectures Notes Univ. Paris XI, 1997.

[2] J.F. Crow and M. Kimura. *An introduction to population genetics theory.* Harper and Row, New York, 1970.

[3] R. Cerf. The dynamics of mutation-selection algorithms with large population sizes. *Ann. Inst. H. Poincaré Probab. Statist.*, **32**, 455-508, 1996.

[4] R. Cerf. Asymptotic Convergence of Genetic Algorithms. *Adv. Applied Probab.*, **30**, 521-550,1998.

[5] D.B. Fogel. *Evolutionary Computation: toward a new philosophy of machine intelligence*, IEEE Press, New York, 1995.

[6] D.B. Fogel ed. *The fossil record.* IEEE Press, Pitsicaway, 1999.

[7] O. François. An evolutionary strategy for global minimization and its Markov chain analysis, *IEEE Transactions on Evolutionary Computation*, **2**, 77-90, 1998.

[8] O. François. Global optimization with exploration/selection procedures and simulated annealing, *Ann. Applied Probab.*, , 2002.

[9] M.I. Freidlin and A.D. Wentzell. Random perturbations of dynamical systems, Springer Verlag, New York, 1984.

[10] B. Hajek. Cooling schedules for optimal annealing, *Math. Oper. Res*, **13** , 311-329, 1988.

[11] S. Karlin. *A first course in stochastic processes.* Academic press, New York, 1968.

[12] Kemeny, L. Snell. *Finite Markov chains*,D. Van Nostrand Company, Princeton,1960.

[13] M. Kimura *The neutral theory of molecular evolution.* Academic Press, New York, 1983.

[14] T. Nagylaki. *Introduction to theoretical population genetics.* Springer Verlag, Berlin, 1992.

[15] A.Trouvé. Cycle decomposition and simulated annealing, *SIAM J. Control Optim.*, **34**, 966-986, 1996.