Survival-extinction phase transition in a bit-string population with mutation

Kathia M. Fehsenfeld,[1,*] Ronald Dickman,[1,†] and Américo T. Bernardes[2,‡]
[1]*Departamento de Física, ICEx, Caixa Postal 702, Universidade Federal de Minas Gerais, 30123-970 Belo Horizonte, Minas Gerais, Brazil*
[2]*Departamento de Física, Universidade Federal de Ouro Preto, 35400-000 Ouro Preto, Minas Gerais, Brazil*
(Received 17 September 2002; published 24 March 2003)

A bit-string model for the evolution of a population of haploid organisms, subject to competition, reproduction with mutation, and selection, is studied, using mean-field theory and Monte Carlo simulations. We show that, depending on environmental flexibility and genetic variability, the model exhibits a phase transition between extinction due to random drift and survival. For weak selection the population attains a neutral regime. The mean-field theory describes the infinite-size limit, while simulations are used to study quasistationary properties.

DOI: 10.1103/PhysRevE.67.031915 PACS number(s): 87.23.Cc, 87.23.Kg, 87.10.+e

I. INTRODUCTION

Many mathematical models have been proposed to describe the evolution of populations [1], focusing on varied aspects, for example, mutation accumulation [2–6] and adaptation [7–14]. Since its introduction in the context of prebiotic evolution, Eigen's model [2] of chemical replicators has attracted increasing interest in the mathematical description of populations subject to natural selection and mutation. As a model of molecular and viral evolution, many authors use it to study competition between replicators of the same species with different production rates [2,3] or different kinds of replicators as, for example, competition between a viral population and the immune system [12]. Recently, the quasispecies model was also used to study the more fundamental problem of stability of different kinds of replicators [14].

A related problem is the development of a simple model capable of describing the response of a population to environmental mutability. Of interest, for example, is the ability of a population to adapt to rapid changes in its environment. In this paper, we propose a model in which a genome is represented as a string of binary symbols (a ''bit-string'') [4,15], subject to mutation and selection. We use the model to study the consequences of variation of the conditions affecting survival, related to environmental flexibility, and the genetic variability of the population. Our main interest is to describe the conditions determining the extinction or survival of the population. While our model (to be defined below) has certain aspects in common with Eigen's model (discrete genomic sequences, genome-dependent reproduction efficiency), an important difference is that here, as is common in population biology studies [8], we allow the population size to fluctuate, even to the point of extinction.

In our model, the population evolves in discrete time with nonoverlapping generations. It consists of haploid organisms defined by their genotype (a bit-string of G positions, or genes). The individuals undergo asexual reproduction, subject to mutation, competition, and selection. Selection is represented though a survival probability that depends on the difference between a genome and a certain *ideal* genome. Varying the parameters in the survival probability, the implied ''fitness landscape'' varies from one having a single sharp peak to one having a broad maximum. Environmental changes can be represented via alteration of this ideal. In the present study, however, the ideal genome is fixed, allowing a systematic analysis of the effect of various other parameters upon survival, so as to provide a benchmark for understanding the effects of a variable environment in future work.

We develop a mean-field theory (MFT) that describes the evolution of an infinite population exactly, since the latter has no spatial structure. We also perform Monte Carlo simulations of the model. The latter are useful for studying fluctuations due to finite population size, which are not captured in the MFT. We determine the survival-extinction phase boundary, and compare the temporal evolution, and the genomic distribution of the population predicted by the MFT against simulation results.

The paper is organized as follows. In Sec. II, we define the model and in Sec. III, we develop the MFT. Section IV describes the Monte Carlo simulation algorithm, while Sec. V reports the MFT and the simulation results. We present our conclusions in Sec. VI.

II. MODEL

We study a model for evolution of a population of haploid individuals defined by their genomes, subject to competition, asexual reproduction with mutation, and selection. In this model, successive generations do not overlap. Each individual is represented by a bit-string of G positions (genes), denoted by the vector $\sigma=(\sigma_1,\sigma_2,\ldots,\sigma_G)$, where $\sigma_i=0$ or 1. The survival probability of an individual in the given environment is measured in relation to a ''model individual'' (or ''ideal genome''), represented by the sequence $\sigma_i=0$, $i=1,\ldots,G$. Each gene in state 1 represents a reduction in survival probability, and carries the same weight, independent of its position i. Thus, the Hamming distance from the ideal genome, given by $H=\Sigma_i\sigma_i$, characterizes an

*Email address: kathia@escelsa.com.br
†Email address: dickman@fisica.ufmg.br
‡Email address: atb@iceb.ufop.br

individual's survival probability. (This type of survival probability has been used in several studies of age-structured populations [9–11].) The survival probability is given by

$$S(H)=\frac{1+e^{B}}{e^{H/G\tau}+e^{B}}, \tag{1}$$

where $S(H)$ is the probability for an individual to survive up to the stage in which she must compete with the rest of the population; individuals that survive the competition stage go on to reproduce offspring, as detailed below. The parameter τ, which plays a role analogous to temperature in equilibrium statistical mechanics, represents environmental flexibility. The parameter B is related to the genetic variability of the population, and represents mutational tolerance, playing a role analogous to the selection factor defined in certain theories of populations, to describe the influence of deleterious mutations on the survival probability [15]. $S(H)=1$ for $H=0$, and decays monotonically with H. We note that for fixed H and B, the survival probability is an increasing function of τ, and that for fixed H and τ, S is an increasing function of B. For small τ and B, $S(H)$ decays rapidly, so that only individuals with H close to zero have an appreciable probability to survive (sharply peaked ''fitness'' landscape). For larger values of B the function exhibits a steplike change from $S\approx 1$ to $S\approx 0$ for $H\simeq BG\tau$, with an inclination $\sim 1/\tau$. The Fermi-like function $S(H)$ was used in a similar manner in the model of Thoms *et al.* [9]. These authors define a death probability $p_d=[e^{\beta(b-a)}+1]^{-1}$, where β is an inverse temperature and $(b-a)$ represents the difference between the typical number of mutations in the population and the number of mutations of the individual.

At reproduction, each organism is replaced by two offspring. The latter are copies of their parent, with a certain number m of mutations. Each position has a probability of λ to mutate (mutations $0\rightarrow 1$ and $1\rightarrow 0$ are considered equally likely), with mutations at different positions constituting independent events. The number of mutations m therefore follows a binomial distribution. The mean number of mutations per reproduction event, λG, is set to unity in this study.

Competition amongst individuals is represented by the familiar Verhulst factor

$$V=1-\frac{N(t)}{N_{max}}, \tag{2}$$

where $N(t)$ is the population at time t and N_{max} is the maximum capacity of the environment. The evolution of the population proceeds by discrete time steps: at each step, the Verhulst factor is applied by selecting at random (independently of H), NV survivors; the survivors go on to reproduce as described above.

III. MEAN-FIELD THEORY

We have developed a mean-field description of the model defined above. For this model, which has no spatial structure, the deterministic mean-field description describes the infinite-size limit ($N_{max}\rightarrow\infty$) exactly. Differences between theory and simulation are due to fluctuations that appear in finite-sized systems, but that are absent in the infinite-size limit.

In the full stochastic description there are 2^G distinct genomes σ, and an integer-valued random variable $N_\sigma(t)\geqslant 0$ for each. Our first step in constructing a simplified description is to reduce the set of variables to $N(H,t)$: the number of individuals with Hamming distance H from the ideal, at time t. Since the model does not distinguish between individuals with the same Hamming distance, the probability distribution at any time $t>0$ will be a function of H only, if it is so at $t=0$. We shall always suppose this to be the case.

In the mean-field theory, the discrete-time evolution of the population may be written as

$$N(H,t+1)=\mathsf{E}[N(H,t+1)|\{N(H,t)\}], \tag{3}$$

where $\{N(H,t)\}$ represents the entire set of population variables at step t. In other words, the population at step $t+1$ is approximated by its *expected value*, given the distribution at step t. (The latter, in turn, is given by the expected distribution, given that for time $t-1$, and so on.) The integer-valued random variables of the exact description are therefore replaced by a set of real-valued, deterministic variables.

Each step of the evolution consists of two stages: (1) death of individuals due to competition for resources (''Verhulst stage'') and (2) reproduction with selection. In the Verhulst stage, the total population size $N=\Sigma_H N(H)$ is evaluated; then each subpopulation is reduced by the same factor, $V=1-N/N_{max}$, yielding the values

$$N'(H)=VN(H), \quad H=0,\ldots,G. \tag{4}$$

Note that the Verhulst stage involves an interaction between individuals [$N'(H)$ is a nonlinear function of all $N(H)$], and that each individual interacts equally with all others in this process.

In the reproduction stage each individual is replaced by a pair of offspring that have, in general, Hamming distances different from those of the parent. We assume independent, equally probable mutations at each site, so that the number of mutations m in a given reproduction event is binomially distributed:

$$P(m)=\binom{G}{m}\lambda^m(1-\lambda)^{G-m}. \tag{5}$$

[Since $G\gg 1$, while the mean number of mutations λG is of order unity, we may approximate $P(m)$ by a Poisson distribution in simulations; we retain the binomial distribution in the MFT analysis.]

Each reproduction event may be represented schematically as $H'\rightarrow H_1,H_2$, where H' denotes the Hamming distance of the parent and H_1 and H_2 those of the offspring. Since $H'\rightarrow H_1$ and $H'\rightarrow H_2$ are independent events (even though they happen simultaneously), it suffices to consider one such, i.e., $H'\rightarrow H$; let $W(H|H')$ represent its probability. If the offspring differs from its parent at exactly m positions, then,

$$\max[0,H'-m] \leqslant H \leqslant \min[H'+m,G].$$

Let $m=m_0+m_1$, with m_0 the number of mutations $0\rightarrow 1$ and m_1 the number of type $1\rightarrow 0$. Each event is characterized by H', m, and m_0. (Evidently, $H=H'+m_0-m_1=H'+2m_0-m$.) The probability of such an event is given by the hypergeometric distribution:

$$p(m_0|m,G,H') = \frac{\binom{G-H'}{m_0}\binom{H'}{m-m_0}}{\binom{G}{m}}. \tag{6}$$

Now using $m_0=(H-H'+m)/2$, we have

$$W(H|H') = (G-H')!H'! \sum_{m=0}^{G} \frac{\lambda^m (1-\lambda)^{G-m}}{\left(\frac{H-H'+m}{2}\right)!\left(\frac{H'-H+m}{2}\right)!\left(G-\frac{H+H'+m}{2}\right)!\left(\frac{H'+H-m}{2}\right)!}. \tag{7}$$

Next we observe that the expected number of *surviving* offspring with Hamming distance H produced by a parent with Hamming distance H' is $\tilde{W}(H|H') \equiv 2S(H)W(H|H')$. Thus the expected number of individuals with Hamming distance H, at step $t+1$ is

$$\mathsf{E}[N(H,t+1)|\{N(H',t)\}] = \sum_{H'=0}^{G} \tilde{W}(H|H')N'(H'), \tag{8}$$

where $N'(H')$ is the distribution just after the Verhulst step. The evolution of the population is found via numerical iteration of Eqs. (4) and (8).

IV. SIMULATION ALGORITHM

We study the evolution of the model population in the Monte Carlo simulations. Initially, $N_0=N_{max}/10$ individuals of $G=128$ bits are generated, each with a random gene sequence, $\boldsymbol{\sigma}=(\sigma_1,\sigma_2,\ldots,\sigma_G)$, where $\sigma_i=0$ or 1 with equal likelihood. The procedure is as follows.

(1) The Verhulst factor $V=1-N(t)/N_{max}$ is evaluated. Then for each individual, a random number s is generated; the individual survives (dies) if $s<V$ ($s>V$).

(2) Each individual reproduces: two copies are created, with possible mutations. The number of mutations m is given by a random integer, chosen from a Poisson distribution with parameter 1. The mutation loci are selected at random.

(3) For each daughter, the Hamming distance H from the ideal is evaluated, and a random number r, uniform on [0,1] is generated. If $r \leqslant S(H)$, the individual survives; otherwise, it dies.

During the simulations, we record the population, average Hamming distance, the average survival probability,

$$\langle S(t)\rangle = \frac{1}{N(t)} \sum_{i=1}^{N(t)} S(H_i), \tag{9}$$

and the *survival rate*, $\mathcal{S}(t) \equiv N(t)/N(t-1)$. (Note that in general $\langle S(t)\rangle < 1$, while $\mathcal{S}(t)$ may, in principle, take any non-negative value, and is unity in the stationary state.) Depending on the parameters τ, B, and N_{max}, the population may survive until a certain maximum time ($t_{max}=30\,000$ steps in the simulations), attaining a quasistationary state, or may go extinct. We record the Hamming distance distribution in the quasistationary state.

V. RESULTS AND DISCUSSION

Depending on the values of B and τ that characterize the survival probability function $S(H)$, Eq. (1), the population either survives or goes extinct. In the mean-field theory this is a sharp transition. In simulations, due to finite population size, fluctuations into the absorbing state (population zero) are to be expected. Indeed, for any *finite* system size the population must eventually go extinct due to random drift, if the process is permitted to continue indefinitely. We adopt $t_{max}=30\,000$ as a convenient maximum time, allowing us to discriminate between survival and extinction, and (in the former case), study quasistationary properties, except very near the transition, where, as noted, the sharp distinction is blurred by fluctuations.

Figure 1 shows the phase boundary between survival and extinction in the B-τ plane, comparing the mean-field prediction against simulations using $N_{max}=10^4, 10^5$, and 5×10^5. As N_{max} is increased, the survival-extinction line found in simulation approaches the MFT prediction, as expected. For small values of τ (a ''hard'' or inflexible environment), survival of the population requires high values of B, the mutational tolerance. The mean-field survival-extinction line of the diagram is obtained by fixing the parameter τ and measuring the stationary population density $\rho=N/N_{max}$ as a function of B. Near the transition, ρ depends linearly on B: $\rho \propto B-B_c(\tau)$, as is normally the case in mean-field descriptions of a continuous phase transition to an absorbing state [16]. The line $B_c(\tau)$ is readily obtained via linear regression to the $\rho(B)$ data near the transition. Note that $B_c=0$ for $\tau>0.192$. For $\tau \ll 1$, on the other hand, $B_c \propto 1/\tau$. (Increasing the mutation probability λ, the phase boundary is displaced upward and to the right, enlarging the extinction region.) Figure 2 is a three-dimensional plot of the population density as a function of B and τ; the extinction region is evident, as is the monotonic growth of ρ with either parameter.

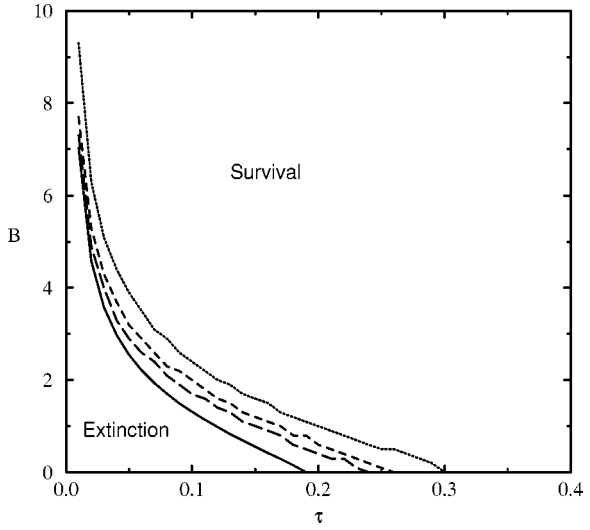

FIG. 1. Survival-extinction phase boundary in the B-τ plane for $\lambda G=1$. The solid line is the MFT prediction; dashed lines represent simulation results for $N_{max}=5\times10^5$, 10^5, and 10^4 (bottom to top).

Figure 3 presents a typical evolution of the population density $\rho(t)$. For B and τ in the survival phase, the population exhibits a rapid initial decay and then evolves to a quasistationary state. Simulation and MFT evolutions are in good agreement, despite fluctuations in the former.

The quasistationary distribution of Hamming distances obtained in simulation is compared in Fig. 4 with the stationary distribution predicted by the MFT. In all cases, the distribution peaks near the mean value $\langle H\rangle$, and has a generally Gaussian appearance. For fixed τ, we observe that $\langle H\rangle$ increases monotonically with B, attaining a *plateau*, if τ is sufficiently large. The plateau value is $\langle H\rangle\simeq64$, i.e., half the genome size, since this corresponds to the largest number of genome sequences. The plateau indicates that the population has attained a neutral regime: individuals have nearly the same survival probability, independent of the Hamming distance, due to weak selection. For fixed B, we observe that $\langle H\rangle$ increases with τ, until attaining $\langle H\rangle=64$. The variance of the distribution behaves similarly. Its saturation value is

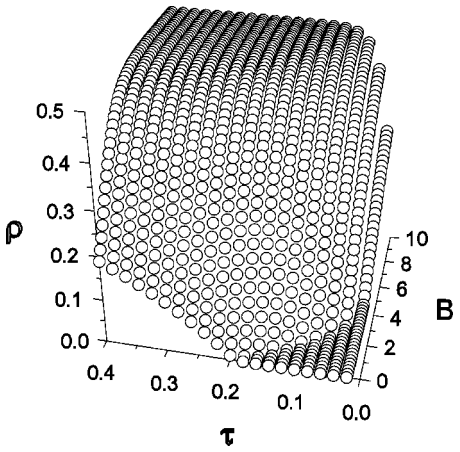

FIG. 2. Population density ρ as a function of B and τ from the MFT. For $\tau\geqslant0.192$, the population survives for any value of B.

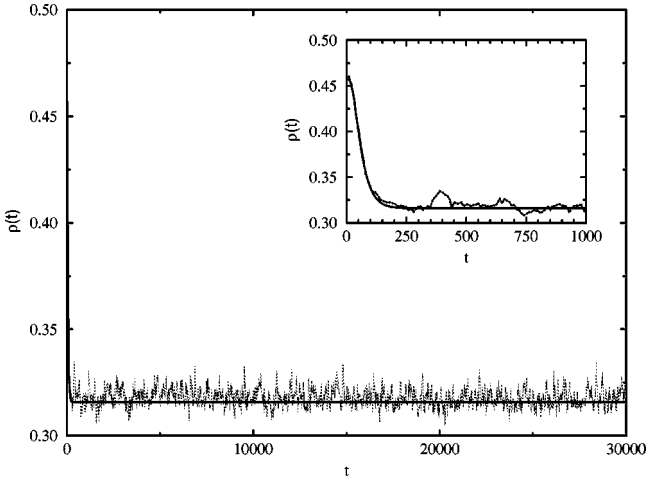

FIG. 3. Time evolution of the population density ρ for $\tau=0.1$ and $B=4$, in the MFT (smooth curve) and simulation ($N_{max}=10^5$).

about 32 the value expected for a binomial distribution with $p=1/2$ and $N=128$, giving a standard deviation $\sigma\simeq5.7$. Figure 5 shows the stationary values of $\langle H\rangle$ and σ_H as functions of B, as predicted by the MFT; simulations yield very similar behavior. In simulations, extinction occurs at larger B values than are predicted by the MFT, due to finite-size effects, as noted above; the difference between simulation and theory diminishes with increasing system size.

VI. SUMMARY

We propose a bit-string model of the evolution of a simple haploid population. Similarly to previous studies [9–11], the model includes the effect of enviromental flexibility and tolerance to genetic differences on the survival probability. Unlike previous works, we employ a survival probability that is a monotonic increasing function of the parameters B and τ that represent tolerance of genetic difference between a given

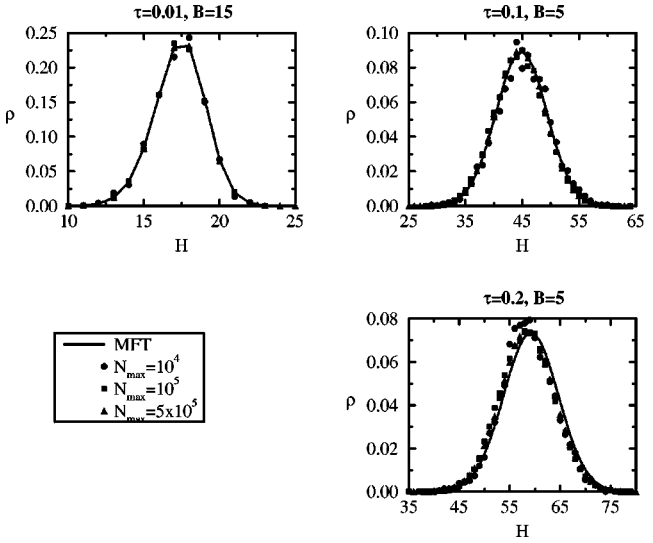

FIG. 4. Stationary Hamming distance distribution for various parameters, as indicated.

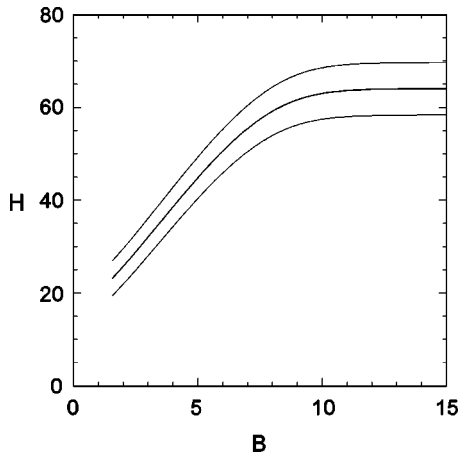

FIG. 5. Dependence of Hamming distance on B for $\tau=0.1$ in the MFT. Central line: mean Hamming distance, $\langle H\rangle$; upper and lower lines represent one standard deviation above or below the mean.

genome and the ideal. The model is studied via computer simulations and the mean-field theory, which are in good agreement.

The model, like many others in population dynamics or epidemic analysis, exhibits a continuous transition between an active phase (survival) and an absorbing one (extinction). We map out the phase boundary in the B-τ plane, and find clear evidence of mean-field-like critical behavior, as in other population models lacking spatial structure [16]. The mean-field description is exact in the infinite-size limit, but provides no information regarding fluctuations. On the other hand, simulations for parameter values in the active phase yield information on the quasistationary state of a finite system ($N_{max}<\infty$). It is also of interest to obtain the *lifetime* of this quasistationary state, or, equivalently, the mean first-passage time to extinction. Such information can, in principle, be obtained from simulations, or from a probabilistic analysis of finite populations starting from the master equation [17]. Given the large number of random variables involved ($G+1$, if we assume that the probability depends only on Hamming distance H), the multivariate Fokker-Planck equation would seem the most convenient tool; theoretical analysis of finite populations is left as subject for future work. The simulation results reported here should prove useful in testing such theories.

Another interesting direction for future study is the response of the population to changes in the environment. Such changes can be represented by variations in the ideal genome (as presented in Refs. [10,11,18]) and/or in the parameters τ, B, λ, and N_{max}. A related question is that of transitions in the genome distribution when two or more ideals (corresponding to distinct, well-adapted types in the ''fitness'' landscape) exist. Studies of these problems using the bit-string model are in progress.

ACKNOWLEDGMENTS

A.T.B. acknowledges the kind hospitality of the Departamento de Física-UFMG. This work was partially supported by the Brazilian Agencies CNPq, FINEP, and FAPEMIG.

[1] B. Drossel, Adv. Phys. **50**, 209 (2001).
[2] M. Eigen, Naturwissenschaften **58**, 461 (1971).
[3] M. Eigen and P. Schuster, Naturwissenschaften **65**, 7 (1978).
[4] T.J.P. Penna, J. Stat. Phys. **78**, 1629 (1995).
[5] A.T. Bernardes, Physica A **230**, 156 (1996).
[6] P.R.A. Campos and J.F. Fontinari, J. Phys. A **32**, L1 (1999).
[7] R. Bürger and M. Lynch, Evolution (Lawrence, Kans.) **49**, 151 (1995).
[8] B. Charlesworth, *Evolution in Age-Structured Populations* (Cambridge University Press, Cambridge, 1994).
[9] J. Thoms, P. Donahue, D. Hunter, and N. Jan, J. Phys. I **5**, 1689 (1995).
[10] S. Cebrat and A. Pekalski, Eur. Phys. J. B **11**, 687 (1999).
[11] A. Pekalski, Physica A **265**, 255 (1999).
[12] C. Kamp, C.O. Wilke, C. Adami, and S. Bornholdt, e-print cond-mat/0209613 (2002).
[13] S. Bonhoeffer and P. Sniegowski, Nature (London) **420**, 367 (2002).
[14] C.P. Ferreira and J.F. Fontanari, Phys. Rev. E **65**, 021902 (2002).
[15] D. Charlesworth, M.T. Morgan, and B. Charlesworth, Genet. Res. **59**, 49 (1992).
[16] J. Marro and R. Dickman, *Nonequilibrium Phase Transitions in Lattice Models* (Cambridge University Press, Cambridge, 1999).
[17] R. Dickman and R. Vidigal, J. Phys. A **35**, 1147 (2002).
[18] C.O. Wilke, C. Ronnewinkel, and T. Martinez, Phys. Rep. **349**, 395 (2001).