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*

(February 1, 2008)

Abstract

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 transtion between extinction and survival. The mean-field theory describes the infinite-size limit, while simulations are used to study quasi-stationary properties.

* Email address: kathia@escelsa.com.br

† Email address: dickman@fisica.ufmg.br

‡ Email address: atb@iceb.ufop.br

I. INTRODUCTION

Many mathematical models have been proposed to describe the evolution of populations, focusing on varied aspects, for example, mutation accumulation [1,2], and adaptation [3,4,5,6,7]. In the first case, how deleterious mutations are passed to offspring, and the consequences for individual growth, are of particular interest. In the second, the principal interest is the influence of different environmental conditions on the population. One goal in this area 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. Penna's bit-string model [1] seems well suited to this purpose.

In this paper, we propose a model for evolution of an adapting population, to study the consequences of variation of 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. The population evolves in discrete time with non-overlapping 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. Environmental changes can be represented via alteration of this ideal. In the present study, however, the ideal genome is fixed, allowing a systematic study of the effect of various other parameters upon survival.

We develop a mean-field (MFT) description, which describes the evolution of an infinite population exactly, since it has no spatial structure. We also perform Monte Carlo simulations for the model. The latter are useful for studying fluctuations due to finite population size, that 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 MFT against simulation results.

The paper is organized as follows. In Sec. II, we define the model and in Sec. III develop the MFT. Sec. IV describes the Monte Carlo simulation algorithm, while Sec. V reports MFT and 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, and 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 $\sigma = (\sigma_1, \sigma_2, ..., \sigma_G)$, where $\sigma_i = 0$ or 1. The fitness of an individual to the environment is measured in relation to a "model individual" (or "ideal genome"), represented by the sequence $\sigma_i = 0$, $i = 1, ..., G$. Each gene in state 1 represents a reduction in fitness, and carries the same weight, independent of its position i. Thus the Hamming distance from the ideal genome, given by $H = \sum_i \sigma_i$, characterizes an individual's fitness (This manner of characterizing fitness has been used in several studies of age-structured populations [5,6,7].) The dependence of fitness on H is through the *survival probability*

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

$S(H)$ 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, while B, which is related to the genetic variability of the population, represents mutational tolerance. $S(H) = 1$ 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. The Fermi-like function $S(H)$ was used in a similar manner in the model of Thoms *et al* [5]. 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 $0 \to 1$ and $1 \to 0$ 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 ($N_{max} \to \infty$) 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 $N_\sigma(t) \geq 0$ 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 so:

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

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'); (2) reproduction/selection. In the Verhulst stage, the total population size $N = \sum_H N(H)$ is evaluated; then each subpopulation is reduced by the same factor, $V = 1 - N/N_{max}$, yielding the values:

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

Note that the Verhulst stage involves an interaction between individuals ($N'(H)$ is a nonlinear function of all of the $N(H)$), 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:

$$P(m) = \begin{pmatrix} G \\ m \end{pmatrix} \lambda^m (1-\lambda)^{G-m} \, . \tag{5}$$

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

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

Let $m = m_0 + m_1$, with m_0 the number of mutations $0 \to 1$ and m_1 the number of type $1 \to 0$. 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:

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

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

i) 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 $s < V$ ($s > V$).

ii) Each individual reproduces: 2 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.

iii) For each daughter, the Hamming distance H from the ideal is evaluated, and a random number r, uniform on [0,1] is generated. If $r \leq S(H)$, 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 nonnegative 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 quasi-stationary state, or may go extinct. We record the Hamming distance distribution in the quasi-stationary 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, 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 quasi-stationary properties, except very near the transition, where, as noted, the sharp distinction is blurred by fluctuations.

Fig. 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 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 [10]. 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.) Fig. 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. 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 quasi-stationary state. Simulation and MFT evolutions are in good agreement, despite fluctuations in the former.

The quasi-stationary distribution of Hamming distances obtained in simulation is compared in Fig. 4 with the stationary distribution predicted by mean-field theory. In all cases, the distribution peaks near the mean value $< H >$, and has a generally Gaussian appearance. For fixed τ, we observe that $< H >$ increases monotonically with B, attaining a *plateau*, if τ is sufficiently large. The plateau value is $< H > \simeq 64$, i.e., half the genome size. For fixed B, we observe that $< H >$ increases with τ, until attaining $< H >= 64$.

The variance of the distribution behaves similarly. Its saturation value is about 32, giving a standard deviation $\sigma \simeq 5.7$. This is not surprising, given that B and τ both represent tolerance of differences from the ideal genome. Fig. 5 shows the stationary values of $< H >$ and σ_H as functions of B, as predicted by MFT; simulations yield very similar behaviour. In simulations, extinction occurs at larger B values than are predicted by 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 a simple haploid population. Similarly to previous studies [5,6,7], 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 genome and the ideal. The model is studied via computer simulations and 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 [10]. 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 quasi-stationary state of a finite system ($N_{max} < \infty$). It is also of interest to obtain the *lifetime* of this quasi-stationary 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 [11]. 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 [6,7]) 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

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

REFERENCES

[1] T. J. P. Penna, J. Stat. Phys. **78**, 1629 (1995).

[2] A. T. Bernardes, Physica A**230**, 156 (1996).

[3] R. Bürger and M. Lynch, Evolution **49**, 151 (1995).

[4] B. Charlesworth, *Evolution in Age-Structured Populations*, (Cambridge University Press, Cambridge, 1994).

[5] J. Thoms, P. Donahue, D. Hunter, N. Jan, J. Phys. I France **5**, 1689 (1995).

[6] S. Cebrat, A. Pekalski, Eur. Phys. J. B **11**, 687 (1999).

[7] A. Pekalski, Physica A **265**, 255 (1999).

[8] N. G. van Kampen, *Stochastic Processes in Physics and Chemistry*, (North-Holland, Amsterdam, 1992).

[9] C. W. Gardiner, *Handbook of Stochastic Methods*, (Springer-Verlag, Berlin, 1990).

[10] J. Marro and R. Dickman, *Nonequilibrium Phase Transitions in Lattice Models*, (Cambridge University Press, Cambridge, 1999).

[11] R. Dickman and R. Vidigal, J. Phys. A**35**, 1147 (2002).

Figure Captions

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 $N_{max} = 5 \times 10^5$, 10^5 and 10^4 (bottom to top).

Fig. 2. Population density ρ as a function of B and τ from MFT. For $\tau \geq 0.192$, 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 MFT (smooth curve) and simulation ($N_{max} = 10^5$).

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

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

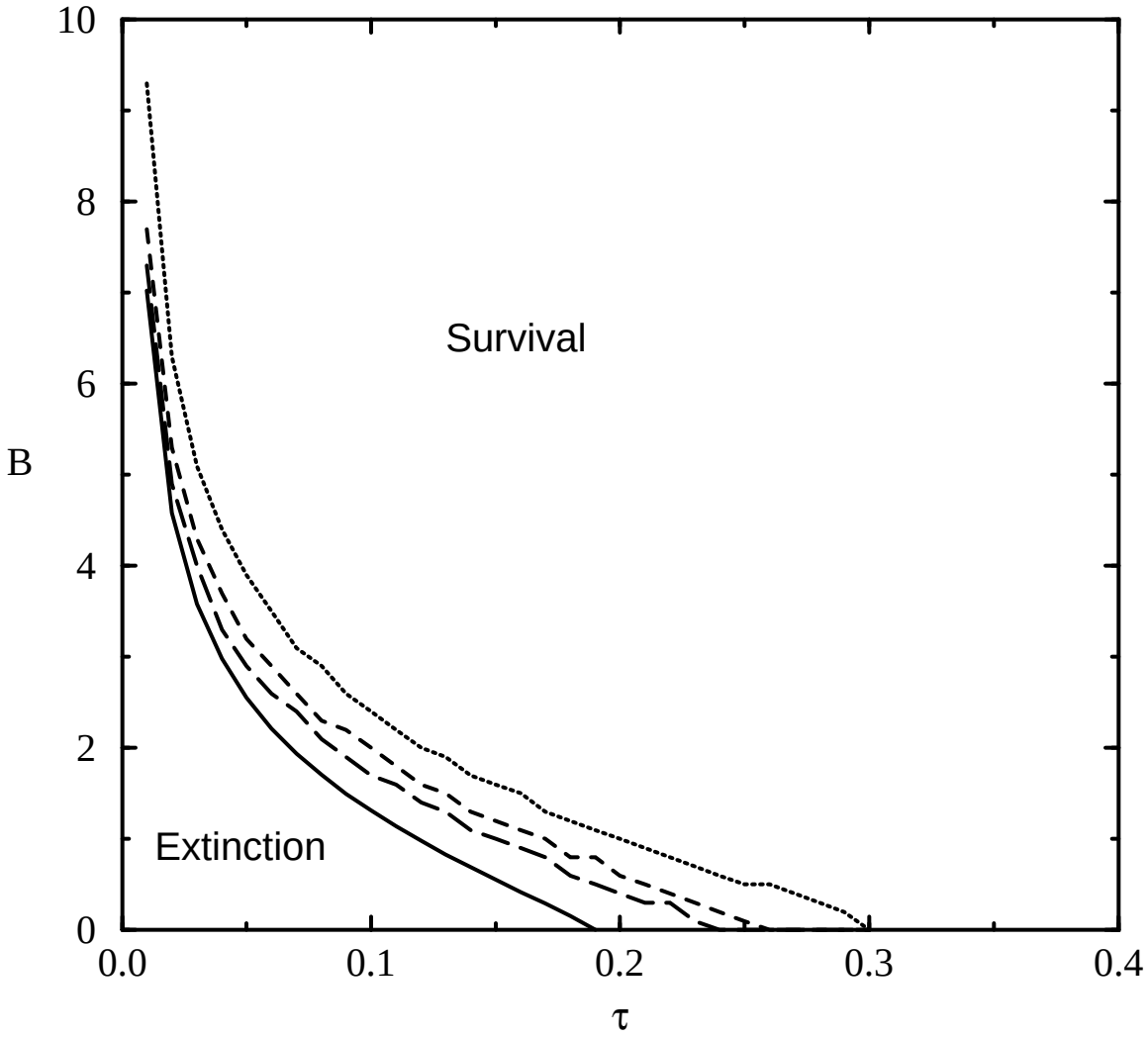


FIG. 1

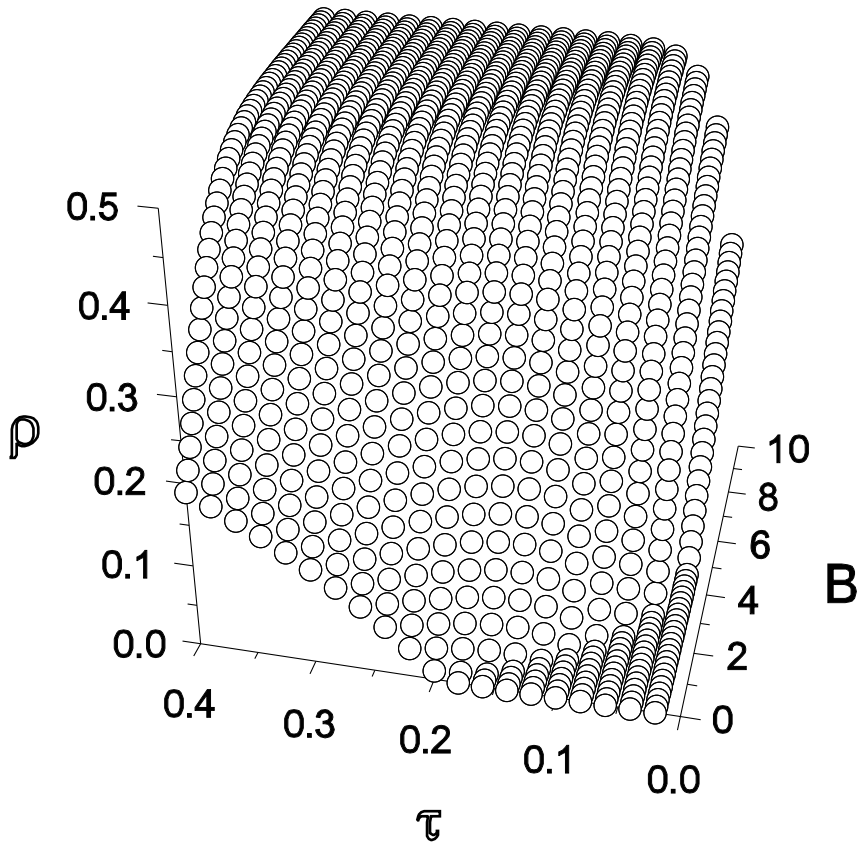


FIG. 2

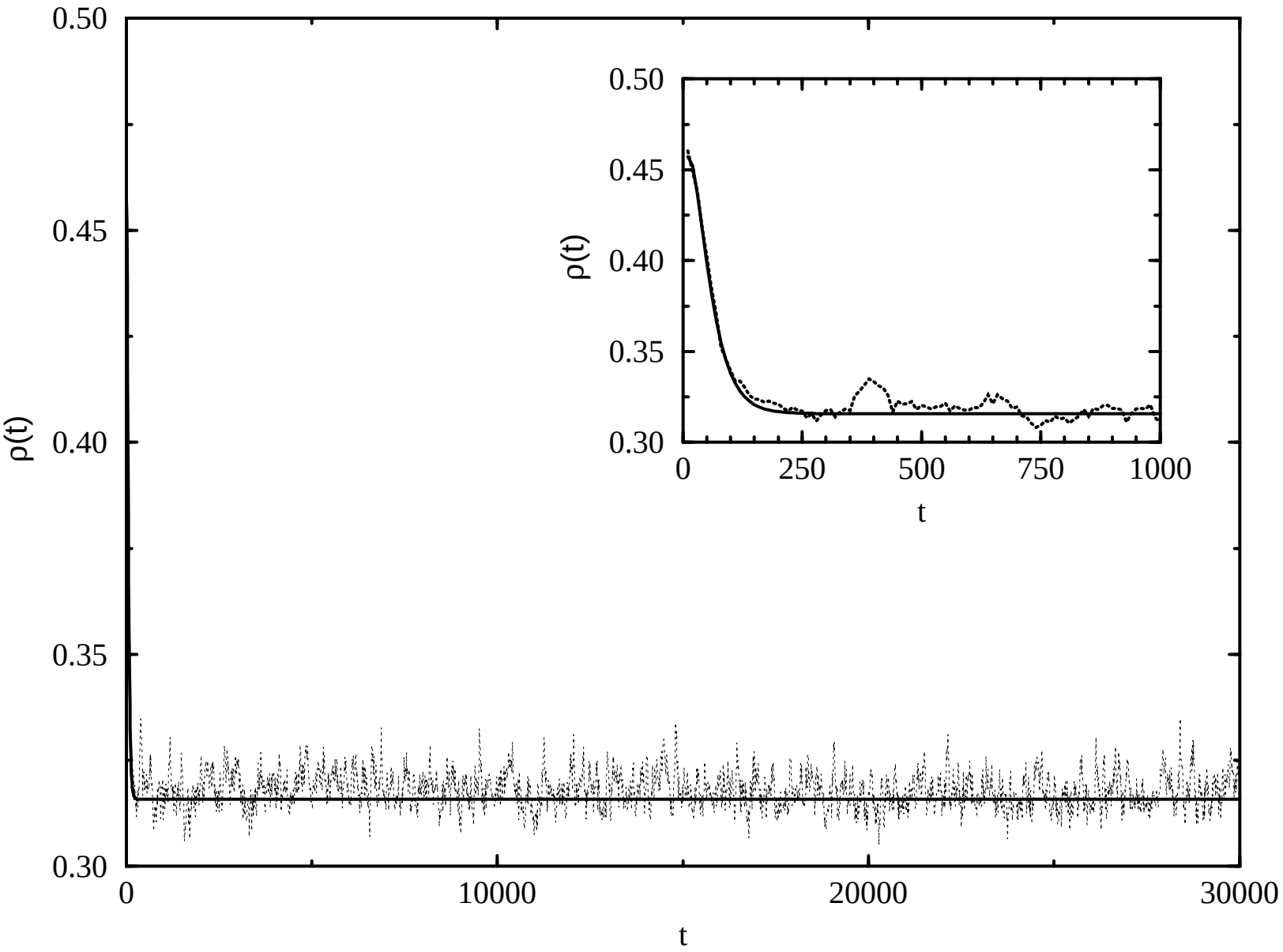
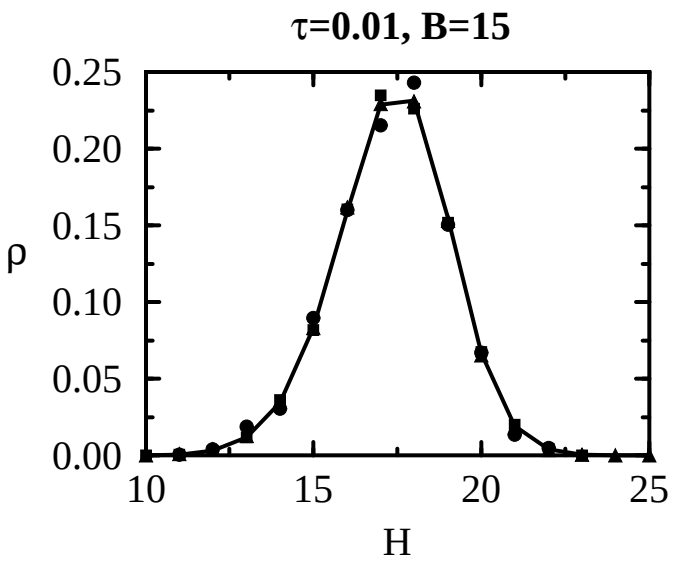
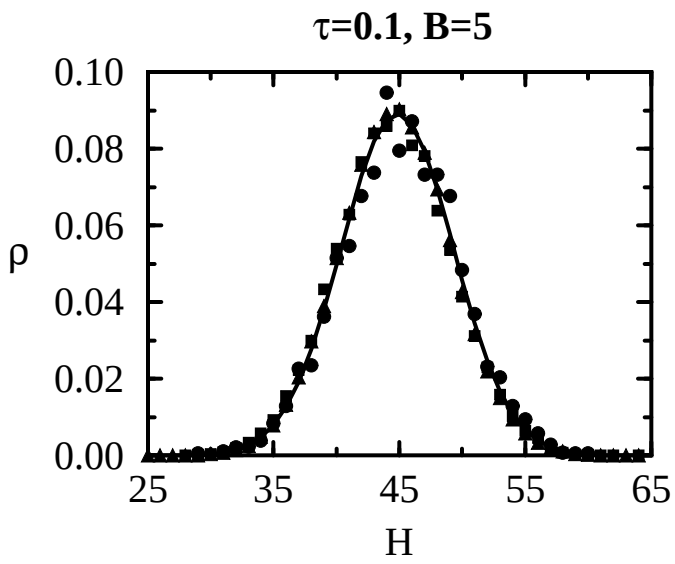
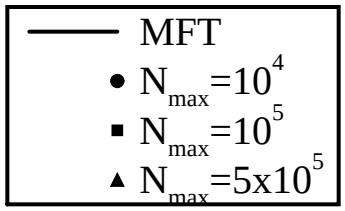

0.50
0.45
0.40
0.35
0.30
ρ(t)
0
10000
20000
30000
t
0.50
0.45
0.40
0.35
0.30
ρ(t)
0
250
500
750
1000
t

FIG 3


τ=0.01, B=15
ρ
H
0.25
0.20
0.15
0.10
0.05
0.00
10
15
20
25


τ=0.1, B=5
ρ
H
0.10
0.08
0.06
0.04
0.02
0.00
25
35
45
55
65


MFT
N_max=10^4
N_max=10^5
N_max=5x10^5

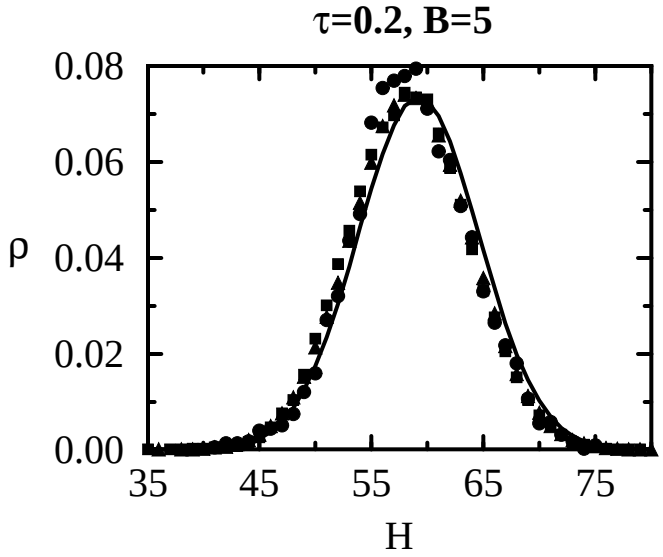

τ=0.2, B=5
ρ
H
0.08
0.06
0.04
0.02
0.00
35
45
55
65
75

FIG 4

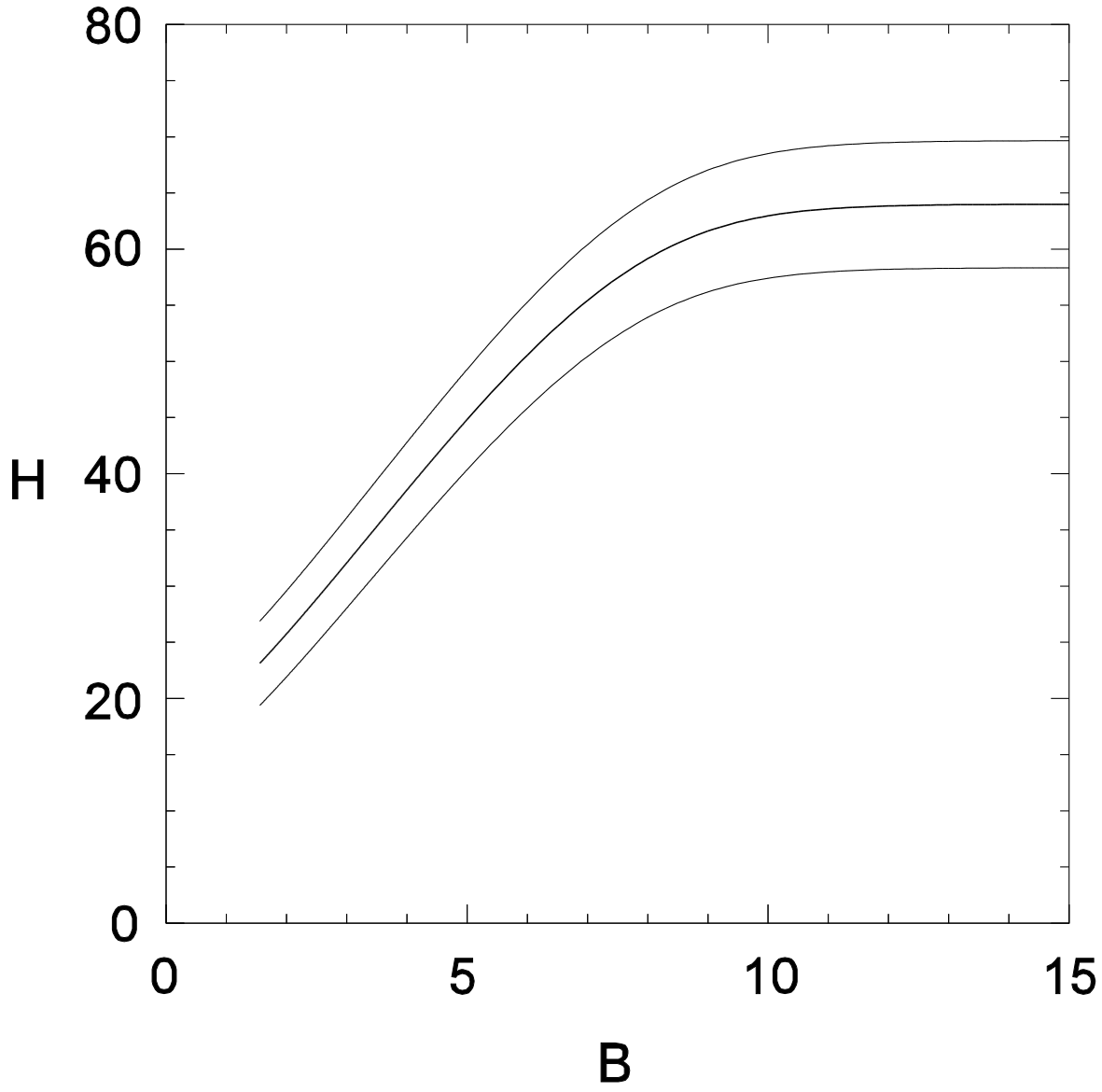


FIG. 5