Critical Exponent of Species-Size Distribution in Evolution

Chris Adami[1], Ryoichi Seki[1,2] and Robel Yirdaw[2]

[1]California Institute of Technology, Pasadena, CA 91125
[2]California State University, Northridge, CA 91330
adami@caltech.edu

Abstract

We analyze the geometry of the species– and genotype-size distribution in evolving and adapting populations of single-stranded self-replicating genomes: here programs in the Avida world. We find that a scale-free distribution (power law) emerges in complex landscapes that achieve a separation of two fundamental time scales: the relaxation time (time for population to return to equilibrium after a perturbation) and the time between mutations that produce fitter genotypes. The latter can be dialed by changing the mutation rate. In the scaling regime, we determine the critical exponent of the distribution of sizes and strengths of avalanches in a system without coevolution, described by first-order phase transitions in single finite niches.

Introduction

Power law distributions in Nature usually signal the absence of a scale in the region where the scaling is observed, and sometimes point to critical dynamics. In Self-Organized-Criticality (SOC) (**??**), for example, power law distributions reveal the dynamics of an unstable critical point, brought about by slow driving and a feed-back mechanism between order parameter and critical parameter. The critical dynamics is usually described within the language of second-order phase transitions in condensed matter systems (**?**), but it can be shown that SOC-type behavior also occurs within a dual description in terms of the Landau-Ginzburg equation as *first-order* transitions (**?**). Indeed, it was shown that a power law distribution of *epoch-lengths*, that is, the time a particular species dominates the dynamics of an adapting population, is explained by a self-organized critical scenario (**?**) that carries the hallmark of first-order phase transitions. Here, we measure the distribution of abundances of *species* and genotypes in an artificial chemistry, (the Avida Artificial Life system **??**) and show that the distribution is scale-free under a broad class of circumstances, confirming the results reported in (**?**). In the next section, we discuss the first-order dynamics in more detail and examine "avalanches of invention" from the point of view of a thermodynamics of information. In Section III, we measure the critical exponent of the power law of genotype abundances in the limit of infinitesimal driving, i.e., infinitesimal mutation rate, and discuss the role of the fitness landscape in shaping the distribution. In Section IV, we repeat the analysis for a higher taxonomic level (that of species) and discuss its relation to the geometric distributions found by **??**. Conclusions about the evolutionary process drawn from the data obtained in this paper are presented in Section V.

Self-Organization in Evolution

The idea that the evolutionary process occurs in spurts, jumps, and bursts rather than gradual, slow and continuous changes has been around for over 75 years (**?**), but has gained prominence as "punctuated equilibrium" through the work of **??**. The general idea is that evolutionary innovations are not bestowed upon an existing species as a whole, gradually, but rather by the emergence of *one* better adapted mutant which, by its superiority, serves as the seed of a new breed that sweeps through an ecological niche and supplants the species previously occupying it. The global dynamics thus has a microscopic origin, as shown experimentally, e.g., in populations of *E. Coli* (**?**).

Such avalanches can be viewed in two apparently contradictory ways. On the one hand we may consider the wave of extinction touching all species that are connected by their ecological relations, a process akin to percolation and therefore suitably described by the language of second-order critical phenomena (**?**). Such a scenario relies on the *coevolution* of species (to build their ecological relations) and successfully describes power-law distributions obtained from the fossil record (**??**). There is, on the other hand, a description in terms of *informational* avalanches that does not require coevolution and leads to the same statistics, as we show here. Rather than contradicting the aforementioned picture (**?**), we believe it to be complementary.

In the following, we set up a scenario in which *information* is viewed as the agent of self-organization in evolving and adapting populations. Information is, in the strict sense of Shannon theory, a measure of correlation between two ensembles: here a population of genomes and the environment it is adapting to. As described elsewhere (**?**), this correlation

grows as the population stores more and more information about the environment via random measurements, implementing a very effective *natural Maxwell demon*. Any time a stochastic event increases the information stored in the population, a wave of extinction removes the less adapted genomes and establishes a new era. Yet, information cannot leave the population as a whole, which therefore may be thought of as protected by a *semi-permeable membrane* for information, the hallmark of the Maxwell demon. Let us consider this dynamics in more detail.

The simple living systems we consider here are populations of self-replicating strings of instructions, coded in an alphabet of dimension $\mathcal{D}$ with variable string length ℓ. The total number of possible strings is exponentially large. Here, we consider the subset of all strings currently in existence in a finite population of size N, harboring N_g different types, where $N_g \ll \mathcal{D}^\ell$. Each *genotype* (particular sequence of instructions) is characterized by its replication rate ϵ_i, which depends on the sequence only, while its survival rate is given by $\epsilon_i/\langle\epsilon\rangle$, in a "stirred-reactor" environment that allows a mean-field picture. This average replication rate $\langle\epsilon\rangle$ characterizes the fitness of the population as a whole, and is given by

$$\langle\epsilon\rangle = \sum_i^{N_g} \frac{n_i}{N}\epsilon_i \ , \tag{1}$$

where n_i is the *occupation number*, or frequency, of genotype i in the population. As N_g is not fixed in time, the average depends on time also, and is to be taken over all genotypes currently living. The total abundance, or size, of a genotype is then

$$s_i = \int_0^\infty n_i(t)\,dt = \int_{T_c}^{T_e} n_i(t)\,dt \ , \tag{2}$$

where T_c is the time of creation of this particular genotype, and T_e the moment of extinction. Before we obtain this distribution in Avida, let us delve further into the statistical description of the extinction events.

At any point in time, the fate of every string in the population is determined by the craftiness of the best adapted member of the population, described by $\epsilon_{\rm best}$. In this simple, finite, world, which does not permit strings to affect other members of the population except by replacing them, not being the best reduces a string to an ephemeral existence. Thus, every string is characterized by a *relative* fitness, or *inferiority*

$$E_i = \epsilon_{\rm best} - \epsilon_i \tag{3}$$

which plays the role of an *energy* variable for strings of information IAL. Naturally, $\langle E\rangle = 0$ characterizes the *ground state*, or vacuum, of the population, and strings with $E_i > 0$ can be viewed as occupying *excited* states, soon to "decay"

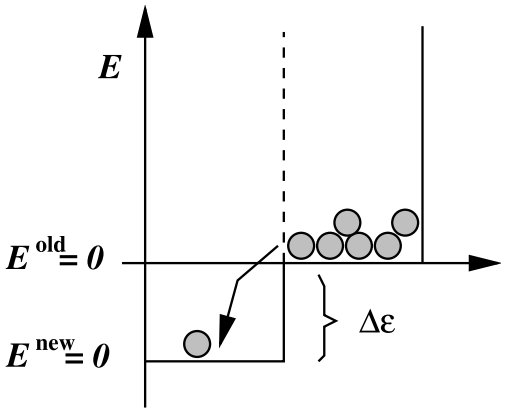

Figure 1: "Energies" (inferiorities) of strings in a first-order phase transition with latent heat $\Delta\epsilon$.

to the ground state (by being replaced by a string with vanishing inferiority). Through such processes, the dynamics of the system tend to minimize the average inferiority of the population, and the fitness landscape of replication rates thus provides a Lyapunov function. Consequently, we are allowed to proceed with our statistical analysis. Imagine a population in equilibrium, at minimal average inferiority as allowed by the "temperature": the rate (or more precisely, the probability) of mutation. Imagine further that a mutation event produces a new genotype, fitter than the others, exploiting the environment in novel ways, replicating faster than all the others. It is thus endowed with a new best replication rate, $\epsilon_{\rm best}^{\rm new}$, larger than the old "best" by an amount $\Delta\epsilon$, and redefining what it means to be inferior. Indeed, all inferiorities must now be *renormalized*: what passed as a ground state ($E = 0$) string before now suddenly finds itself in an excited state. The seed of a new generation has been sown, a phase transition must occur. In the picture just described, this is a first-order phase transition with latent heat $\Delta\epsilon$ (see Fig.**??**), starting at the "nucleation" point, and leading to an expanding *bubble* of "new phase".

This bubble expands with a speed given by the Fisher velocity

$$v \sim \sqrt{D\Delta\epsilon} \ , \tag{4}$$

where D is the diffusion coefficient (of information) in this medium, until the entire population has been converted (**?**). This marks the end of the phase transition, as the population returns to equilibrium via mutations acting on the new species, creating new diversity and restoring the *entropy* of the population to its previous value. This prepares the stage for a new avalanche, as only an equilibrated population is vulnerable to even the smallest perturbation. The system has returned to a critical point, driven by mutations, self-organized by information.

Thus we see how a first-order scenario, without coevolution, can lead to self-organized and critical dynamics. It

takes place within a single, finite, ecological niche, and thus does not contradict the dynamics taking place for populations that span many niches. Rather, we must conclude that the descriptions complement each other, from the single-niche level to the ecological web. Let us now take a closer look at the statistics of avalanches in this model, i.e., at the distribution of genotype sizes.

Exponents and Power Laws

In this particular system avalanche size can be approximated by the size s of the genotype that gave rise to it, Eq. (**??**). We shall measure the distribution of these sizes $P(s)$ in the Artificial Life system Avida, which implements a population of self-replicating computer programs written in a simple machine language-like instruction set of $\mathcal{D} = 24$ instructions, with programs of varying sequence length. In the course of self-replication, these programs produce mutant off-spring because the `copy` instruction they use is flawed at a rate R errors per instruction copied, and adapt to an environment in which the performance of *logical* computations on externally provided numbers is akin to the catalysis of chemical reactions (**?**). In this *artificial chemistry* therefore, successful computations accelerate the metabolism (i.e., the CPU) of those strings that carry the *gene* (code) necessary to perform the trick, and any program discovering a new trick is the seed of another avalanche.

Avida is not a stirred-reactor environment (although one can be simulated). Rather, the programs live on a two-dimensional grid, each program occupying one site. The size of the grid is finite, and chosen in these experiments to be small enough that avalanches are generally over before a new one starts. As is well-known, this is the condition *sine qua non* for the observation of SOC behavior, a separation of time scales which implies that the system is driven at infinitesimal rates.

Let τ denote the average duration of an avalanche. Then, a separation of time scales occurs if the average time between the production of new seeds of avalanches is much larger than τ. New seeds, in turn, are produced with a frequency $\langle\epsilon\rangle P$, where $\langle\epsilon\rangle$ is again the average replication rate, and P is the mutation probability (per replication period) for an average sequence of length ℓ,

$$P = 1 - (1 - R)^{\ell} \, . \tag{5}$$

For small enough R and not too large ℓ (so that the product $R\ell$ is smaller than unity) we can approximate $P \approx R\ell$, and infinitesimal driving occurs in the limit

$$\langle\epsilon\rangle R\ell \ll \frac{1}{\tau} \, . \tag{6}$$

Furthermore

$$\tau \sim \frac{L}{v} \tag{7}$$

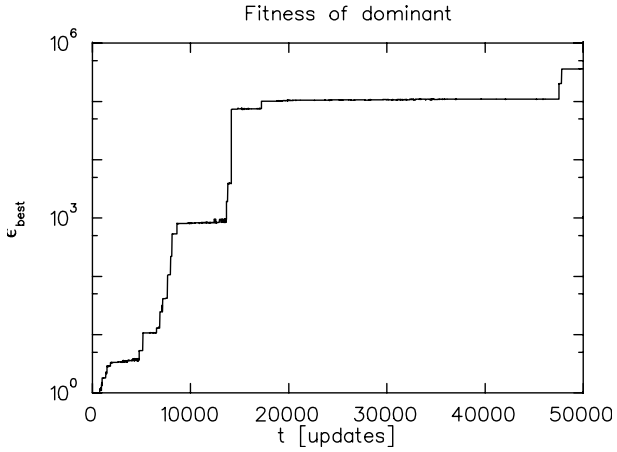

Figure 2: Fitness of the dominant genotype in the population, $\epsilon_{\rm best}$ as a function of time (in updates).

with L the diameter of the system and v a typical Fisher velocity. The fastest waves are those for which the latent heat is of the order of the new fitness, i.e., $\Delta\epsilon \sim \epsilon$, in which case $v \approx \epsilon$ (because $D \sim \epsilon$ in Eq. (**??**), **?**), and a separation of time scales is assured whenever

$$\frac{1}{R\ell} \gg L \, , \tag{8}$$

that is, in the limit of vanishing mutation rate or small population sizes. For the $L = 60$ system used here, this condition is obeyed (for the fastest waves) only for the smallest mutation rate tested and sequence lengths of the order of the ancestor.

In the following, we keep the population size constant (a 60×60 grid) and vary the mutation rate. From the previous arguments, we expect true scale-free dynamics only to appear in the limit of small mutation rates. As in this limit avalanches occur less and less frequently, this is also the limit where data are increasingly difficult to obtain, and other finite size effects can come into play. We shall try to isolate the scale-free regime by fitting the distribution to a power law

$$P(s) \sim s^{-D(R)} \tag{9}$$

and monitor the behavior of D from low to high mutation rates.

In Fig. **??**, we display a typical history of $\epsilon_{\rm best}$, i.e., the fitness of the dominant genotype.[1] Note the "staircase" structure of the curve reflecting the "punctuated" dynamics, where each step reflects a new avalanche and concurrently an extinction event. Staircases very much like these are also observed in adapting populations of *E. Coli* (**?**).

[1] As the replication rate ϵ is exponential in the bonus obtained for a successful computation, $\epsilon_{\rm best}$ increases exponentially with time.

As touched upon earlier, the Avida world represents an environment replete with information, which we encode by providing bonuses for performing logical computations on externally provided (random) numbers. The computations rewarded usually involve two inputs A and B, are finite in number and listed in Table 1. At the end of a typical run (such as Fig. **??**) the population of programs is usually proficient in almost all tasks for which bonuses are given out, and the genome length has grown to several multiples of the initial size to accommodate the acquired information.

Name	Result	Bonus b_i	Difficulty
Echo	I/O	1	–
Not	$\neg A$	2	1
Nand	$\neg(A \wedge B)$	2	1
Not Or	$\neg A \vee B$	3	2
And	$A \wedge B$	3	2
Or	$A \vee B$	4	3
And Not	$A \wedge \neg B$	4	3
Nor	$\neg(A \vee B)$	5	4
Xor	A xor B	6	4
Equals	$\neg(A$ xor $B)$	6	4

Table 1: Logical calculations on random inputs A and B rewarded, bonuses, and difficulty (in minimum number of `nand` instructions required). Bonuses b_i increase the speed of a CPU by a factor $\nu_i = 1 + 2^{b_i - 3}$.

Because the amount of information stored in the landscape is finite, adaptation, and the associated avalanches, must stop when the population has exhausted the landscape. However, we shall see that even a 'flat' landscape (on which evolution is essentially neutral after the sequence has optimized its replicative strategy) gives rise to a power law of genotype sizes, as long as the programs do not harbor an excessive amount of "junk" instructions. A typical abundance distribution (for the run depicted in Fig. **??**) is shown in Fig. **??**.

As mentioned earlier, we can also turn *off* all bonuses listed in Tab. 1, in which case fitness is related to replicative abilities only. Still, avalanches occur (within the first 50,000 updates monitored) due to minute improvements in fitness, but the length of the genomes typically stays in the range of the ancestor, a program of length 31 instructions. We expect a change of dynamics once the "true" maximum of the local fitness landscape is reached, however, we did not reach this regime in the experiments presented here. The distribution of genotype sizes for the flat landscape is depicted in Fig. **??**.

Clearly then, even such landscapes (flat with respect to all other activities except replication) are not neutral. Indeed, it is known that neutral evolution, where the chance for a genotype to increase or decrease in number is even, leads to a power law in the abundance distribution with exponent $D = 1.5$ (**?**).

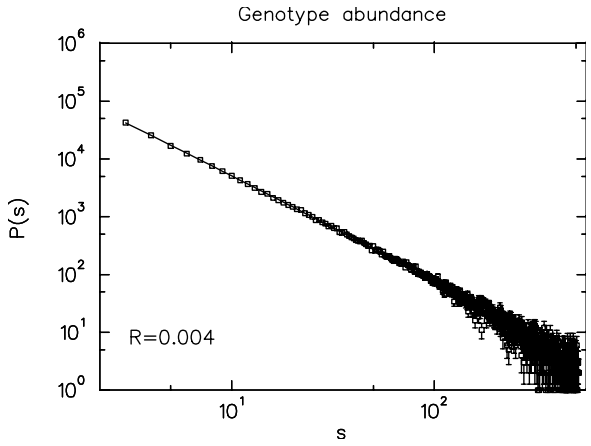

Figure 3: Distribution of genotypes sizes $P(s)$ fitted to a power law (solid line) at mutation rate $R = 0.004$.

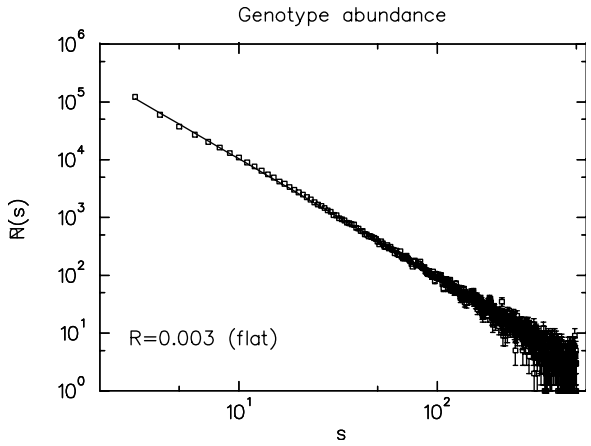

Figure 4: Distribution of genotypes sizes $P(s)$ for a landscape devoid of the bonuses listed in Tab. 1, at mutation rate $R = 0.003$.

In order to test the dependence of the fitted exponent $D(R)$ [Eq. (**??**)] on the mutation rate, we conduct a set of experiments at varying copy-mutation rates from 0.5×10^{-3} to 10×10^{-3} and take data for 50,000 updates. Again, a "best" genotype is not reached after this time, and we must assume that avalanches were still occurring at the end of these runs. Furthermore, in some runs we find that a genotype comes to dominate the population (usually after most 'genes' have been discovered) which carries an unusual amount of junk instructions. As mentioned earlier, such species produce a distribution that is exponentially suppressed at large genotype sizes (data not shown). To avoid contamination from such species, we stop recording genotypes after a plateau of fitness was reached, i.e., if the population had discovered most of the bonuses. Furthermore, in order to minimize finite size effects on the determination of the critical

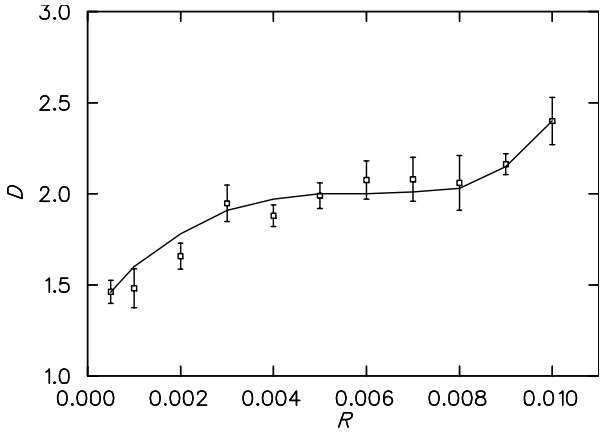

Figure 5: Fitted exponent of power law for 34 runs at mutation rates between $R = 0.0005$ and $R = 0.01$ copy errors per instruction copied. The error bars reflect the standard deviation across the sample of runs taken at each mutation rate. The solid line is to guide the eye only.

exponent, we excluded from this fit all genotype abundances larger than 15, i.e., we only fitted the smallest abundances. Indeed, at larger mutation rates the higher abundances are contaminated by a pile-up effect due to the toroidal geometry, while at lower mutation rates a scale appears to enter which prevents scale-free behavior. We have not, as yet, been able to determine the origin of this scale.

In the results reported here, we show the dependence of the fitted exponent D as a function of the mutation rate R used in the run, which, however, is a good measure of the mutation probability P only at small R and if the sequence length is not excessive. As a consequence, data points at large R, as well as runs where an excessive sequence length developed, carry a systematic error.

Acknowledgements

This work was supported by NSF grant No. PHY-9723972.