Anomalies and the Eve effect in the asexual Penna model

T. Preece and Y. Mao
School of Physics and Astronomy, University of Nottingham, University Park, Nottingham NG7 2RD, United Kingdom
(Received 16 June 2006; revised manuscript received 5 September 2006; published 21 November 2006)

The Penna model of evolutionary ageing is an influential model of mutation accumulation and selection, where an individual's genomic information is represented by a binary bit string. One key parameter of the model is the death threshold, T, the number of diseases any particular individual is able to endure. We show, by combined computer simulations and analytical formulation, that certain anomalies emerge in the asexual Penna model for $T>1$, which may lead to the so-called Eve effect. We characterize these anomalies and their associated demographic distributions. We argue that this anomaly is similar in nature to the well known first-passage problem.

DOI: 10.1103/PhysRevE.74.051915 PACS number(s): 87.23.−n, 87.10.+e

I. INTRODUCTION

In 1995, Penna [1] proposed a binary bit-string computational model for the process of evolutionary ageing. The model is deceptively simple to construct, and yet it captures some key features of evolution, namely, mutation accumulation and selection. Indeed, the basic Penna model provides a useful foundation upon which other effects could be added and studied [2]. As a result, the Penna model has acquired a considerable popularity, and over 170 published citations of the original 1995 article can be found at the time of this writing.

The idea that the natural selection, and therefore the survival of the fittest, seemingly contradicts the detrimental behavior of ageing and the general decline of an organism's capability [3]. The resolution of this conflict lies in the occurrence of mutations. It is now generally accepted that ageing is regulated by specific genes, as originally proposed by Medawar [4], and their effects depend on the reproductive life cycle of the individual organism as well as random mutations that occur [5]. The Penna model provides a means to model this delicate interplay during the evolution of an age-structured population under the influence of age-specific harmful mutations [2].

The original Penna model is designed for computer simulations, and therefore is discrete in nature. Time steps are counted by an integer and an organism's genome represented by a binary bit string. Each 0 on the bit string represents a healthy site, and each 1 a harmful one. The location of the harmful sites on the bit-string, x, indicate the ages at which the organism suffers the harmful effect (a disease). Having suffered T diseases an organism gives up and dies. The bit string is of course finite in length (usually 32 or 64 bits as dictated by the available 32-bit and 64-bit computer processors) and each newborn inherits the parental string, with extra mutations introduced into its bit string. Normally, mutations occur infrequently and multiple mutations are rare. Mutations are also considered as always harmful (turns a 0 into a 1), since harmful mutations vastly outnumber the helpful ones in nature [5]. This assumption may be relaxed to allow a small rate of positive mutation which, provided it is much smaller than the negative mutation rate, does not qualitatively alter the overall demography [6]. Therefore, we confine our discussion here to the case of only single harmful mutations at births.

The large number of Penna simulations, published in the wake of the original Penna model, were not always consistent. For example, Malarz [7] investigated the effects of different bit string lengths on the Penna model. He inquired as to whether large bit strings were required or whether one could expect, after appropriate scaling of other parameters, the same results for different genome lengths. Investigating the effects of string length through simulations, Malarz was unable to find universal agreement in the scaling of the Penna model results. In other words, simulations with different bit string lengths could produce different results, casting doubts over the universal applicability of the Penna model. However, later analytical works [8–10] proved conclusively that universal scaling relations do exist for the Penna model.

In this paper, we show, by combined computer simulations and analytical formulation, that certain anomalies emerge in the asexual Penna model for $T>1$, which lead to the so-called Eve effect [11–13]. We characterize these anomalies and their associated demographic distributions. We suggest that these anomalies may be responsible for some inconsistencies in the earlier Penna simulations. Furthermore, any future Penna simulations must pay special attention to these effects.

II. ANOMALY IN THE PENNA MODEL

The asexual Penna model has been formulated and solved [14,15]. The case of $T=1$ gives straightforward agreement between computer simulation and analytical solution. For $T>1$, the analytical solution can be obtained through an ansatz. We found agreement between this analytical solution and computer simulations provided that either the simulation size is very large or the simulation period is relatively short. With moderate parameters however, computer simulation can frequently lead to a host of distinct steady states. Those are indicative of certain anomalies, which we will analyze here. For the purpose of clarity, we will concentrate on the case of $T=2$. For higher threshold values of T, the anomalies we discuss remain qualitatively the same.

For $T=2$ case, an individual may survive two diseases, and therefore the location of the first two deleterious bits, l_1 and l (with $l>l_1$), are of interest. The second bit at l marks the death of the individual and thus l is referred to as the genetic lifespan. A population can be characterized in terms

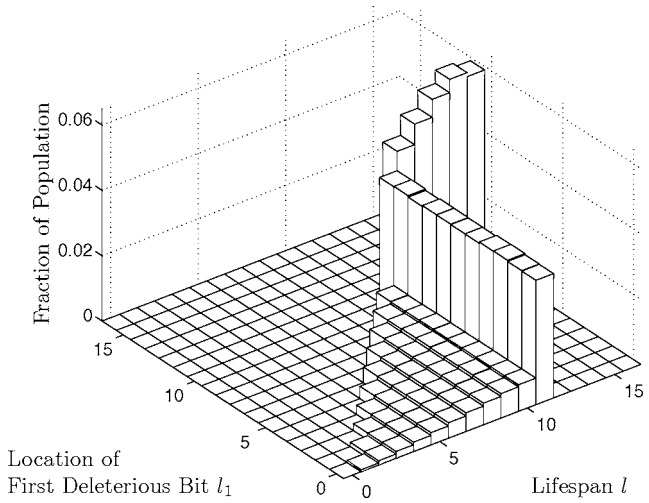

FIG. 1. Simulated population $n(l_1,l)$ versus the location of the first deleterious bit l_1 and the genetic string length l.

of the population distribution function $n(l_1,l)$, defined as the number of individuals with their first two deleterious bits located at l_1 and l, subject to normalization.

Figure 1 presents a fairly typical scenario of $n(l_1,l)$ from a computer simulation of a population after reaching steady state. Simulation parameters are: mutation rate $\beta=1/30$; birth rate is moderated via a Verhulst factor

$$b = b_0\left(1 - \frac{N}{N_{max}}\right),$$

where $b_0=1$, the maximum population $N_{max}=10^7$ and N is the actual population determined at each simulation step; a population reducing Verhulst factor is not introduced though in principle possible; finally $l_{max}=16$ represents the maximum genetic lifespan of the population. These are quite typical simulations which take minutes to hours on a Pentium 4 computer.

A notable feature in Fig. 1 is the formation of a ridge at $l_s=11$. If we sum $n(l_1,l)$ over l_1, we obtain the population as a function of the genetic lifespan l,

$$n(l) = \sum_{l_1} n(l_1,l), \tag{1}$$

then the aforesaid ridge gives rise to a spike in the corresponding $n(l)$ plot, as presented in Fig. 2 (all simulation parameters for Fig. 2 are the same as for Fig. 1). The spike is located at the position of the ridge, $l_s=11$. This is in stark contrast to the prediction of the ansatz which gives rise to a smooth curve for $n(l)$, shown to be valid for a very large population [14]. Clearly, the Penna model contains a potential anomaly that merits investigation.

Further simulations inform us that the spike location l_s appears to be developed randomly for repeated simulations with the same set of initial conditions. Once the spike forms, its location l_s becomes fixed. Furthermore, the population $n(l_1,l)$ dies out completely for the case $l_1>l_s$, and the case

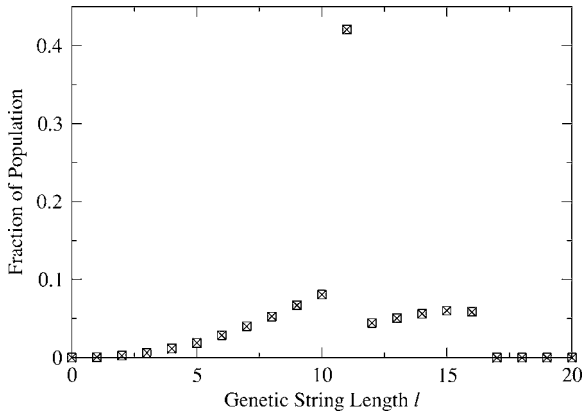

FIG. 2. Population $n(l)$ versus the genetic string length l, from simulation and analytical solution, showing excellent agreement. Squares: simulation; crosses: analytical solution.

$l_1<l_s$ and $l>l_s$. Guided by these observations, we formulate our analytical approach step by step for $l>l_s$, $l=l_s$, $l<l_s$, as follows.

For $l>l_s$. Since population with $l_1>l_s$ dies out, those organisms with $l_1=l_s$ obey the following evolution equation:

$$\frac{dn(l_s,l)}{dt} = be^{-\beta(l-1)}n(l_s,l) - \frac{n(l_s,l)}{l} + bme^{-\beta(l-1)}\sum_{l'=l+1}^{l_{max}} n(l_s,l') \tag{2}$$

where b and β are rates for birth and mutation, respectively; m is the probability of mutation at a given site $m=1-e^{-\beta}$; and l_{max} is the maximum l of the population. The first term on the right-hand side results from mutation free births, the second term from death, and the third term from mutated births. This formulation is similar to the $T=1$ case of the standard Penna model [14]. For steady state, the time derivative d/dt goes to zero, and our evolution equation can be solved exactly by the recursion relation:

$$\frac{n(l_s,l+1)}{n(l_s,l)} = \frac{l+1}{l}\frac{e^{\beta(l-1)} - bl}{e^{\beta l} - b(l+1)e^{-\beta}}. \tag{3}$$

Thus, given the population $n(l_s,l_{max})$, we can determine all $n(l_s,l)$ for $l_s<l<l_{max}$. The model is linear and scalable, so the distribution function n can be normalized later.

For $l=l_s$. The mutated births into $l=l_s$ can only come from those $n(l_s,l)$ with $l_s<l\leqslant l_{max}$, since only single negative mutations are considered. And the evolution equation for population $n(l_1,l_s)$ reads:

$$\frac{dn(l_1,l_s)}{dt} = be^{-\beta(l_s-1)}n(l_1,l_s) - \frac{n(l_1,l_s)}{l_s} + bme^{-\beta(l_s-1)}\sum_{l'=l_s+1}^{l_{max}} n(l_s,l'). \tag{4}$$

Again setting the time derivative to zero for steady state, we obtain the solution

$$n(l_1,l_s) = \frac{bml_s}{e^{\beta(l_s-1)} - bl_s} \sum_{l'=l_s+1}^{l_{max}} n(l_s,l'), \tag{5}$$

where the sum is given by the solution $n(l_s,l)$ from the previous $l>l_s$ case. Note that the right-hand side does not contain l_1, thus our analysis predicts that $n(l_1,l_s)$ is independent of l_1. In other words, the ridge in Fig. 1 should be level, a prediction confirmed by simulations.

For $l<l_s$. Population $n(l_1,l)$ can be enhanced by mutated births from either $n(l_1,l')$ or $n(l,l')$ with $l'>l$, and so we have

$$\frac{dn(l_1,l)}{dt} = be^{-\beta(l-1)}n(l_1,l) - \frac{n(l_1,l)}{l} + bme^{-\beta(l-1)} \times \sum_{l'=l+1}^{l_s} [n(l_1,l') + n(l,l')]. \tag{6}$$

Noting that this equation, in fact, remains true for all $l_1<l$, we deduce that the solution $n(l_1,l)$ will be independent of l_1. This deduction is confirmed by the simulation results shown in Fig. 1, and therefore, Eq. (6) can be solved analytically yielding the following recursion relation:

$$\frac{n(l_1,l+1)}{n(l_1,l)} = \frac{l+1}{l} \frac{e^{\beta(l-1)} - bl}{e^{\beta l} - b(l+1)(2e^{-\beta} - 1)}. \tag{7}$$

Combining the above three stages of analysis, we can calculate the entire distribution of population, provided l_s is known. Since l_s emerges randomly in simulations, it needs to be passed to our analytical formulation in order to match the particular set of simulation data. Finally, normalization can be applied so that $n(l_1,l)$ represents the fractional population so that it becomes independent of the simulated population size. Our analytical results have been plotted in Fig. 2, together with the data from the simulation. The excellent agreement validates the theoretical formulation and the mechanisms it reveals of the population dynamics in the Penna model.

III. DISCUSSION

The standard Penna simulations only consider a possible single negative mutation at birth as the dominant mutation mechanism. This in turn limits the interaction between different genotypes. For a steady-state to exist there must be a longest lived subpopulation which is self-sustaining, i.e., not reliant on mutated births. No other shorter-lived subpopulation can be self-sustaining if the population is to remain bounded, as shorter-lived organisms can always be created by mutated copies of longer-lived ones. For the longest-lived subpopulation to be self-sustaining, each organism must produce on average one perfect copy of itself during its lifetime. All other populations, with $l<l_{max}$, gain from mutated births of the longest lived, so unmutated birth per individual must, on average, be less than unity.

In simulations, birth rate is modified by the so-called Verhulst factor, and when the steady state is reached the birth rate would be such that the organism with $l=l_{max}$ produces on average one mutation-free copy of itself during its lifetime. For $T=1$, there is no ambiguity to the longest living genotype $n(l_{max})$, but for $T=2$ however, it is possible for two individuals to have different locations for their first deleterious bit while having the same $l=l_{max}$ and the same lifespan. Moreover, if $l_1 \neq l_1'$, a subpopulation $n(l_1,l_{max})$ cannot give rise to $n(l_1',l_{max})$ through a single negative mutation. Therefore, the genotype (l_1,l_{max}) would evolve somewhat independently from (l_1',l_{max}). In fact, the evolution equation (2) is identical to the earlier ansatz solution for $T=2$ [14], only the presence of l_s provides alternative population configurations.

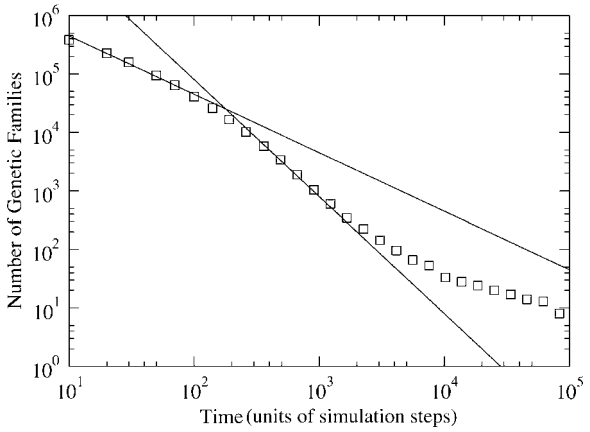

FIG. 3. The decline of the number of distinct genetic families versus time, leading to the Eve effect. The two straight lines have slopes −1 and −2.

In the case of $T=1$, extinction due to the well known Muller's ratchet is alleviated by the Verhulst factor which increase the birth rate when the total population drops. But for $T=2$, all subpopulation $n(l_1,l_{max})$ for different l_1's fluctuate due to the stochastic nature of the simulation, much like a collection of diffusing particles under an overall constraint due to the Verhulst factor. But the Verhulst factor only acts through the total population, and each subpopulation can easily suffer from the ratchet effect as another subpopulation can grow to make up the total population. When a subpopulation $n(l_1,l_{max})$ ventures close to extinction, it receives little help toward a recovery, and the closer it is to extinction the more vulnerable it becomes. Therefore, given enough time, the system would eventually settle into one of subpopulations which gives rise to the spiked states observed in the simulations. This is akin to the first-passage problem of multiple diffusive particles [16]. Figure 3 plots the number of distinct genetic families versus time, showing a steady decline, and leading to the Eve effect. This is in broad agreement with previous results [12] reporting potential scaling regimes of −1 and −2. However, it should be noted that the scaling does not persist over a significant range as is the case in Ref. [12], and the early time behavior is particularly sensitive to initial conditions.

Figure 4 gives a histogram of time taken to reach a spiked population configuration from 1600 repeated simulations under the same initial conditions. The simulation parameters are $\beta=1/30$, $b_0=1$, $l_{max}=16$, and squares with $N_{max}=2.25\times10^4$, crosses with $N_{max}=4.5\times10^4$. The resulting histogram is very similar to the survival time distribution, the so-called

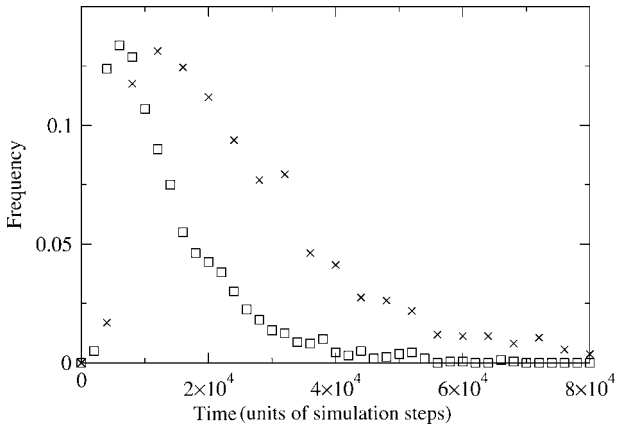

FIG. 4. Histogram of normalized frequency vs time to complete spike formation. Simulation parameters are $\beta=1/30$, $b_0=1$, $l_{max}=16$, and squares with $N_{max}=2.25\times10^4$, crosses with $N_{max}=4.5\times10^4$.

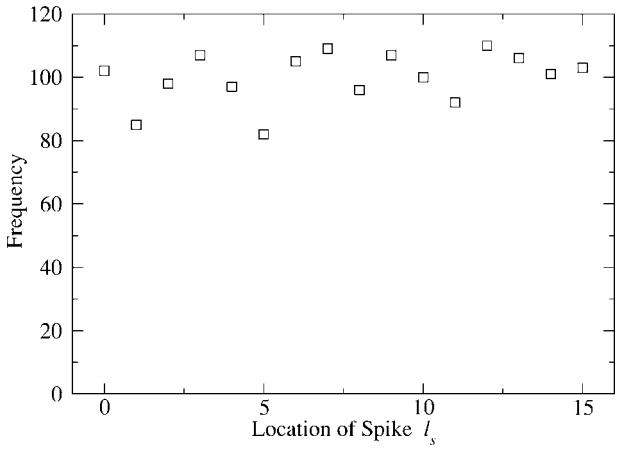

FIG. 5. Distribution of the spike location l_s in 1600 repeated sets of simulations.

Smirnov density, of the first-passage problem [16]. Larger population gives rise to a larger survival time, in accordance with the first-passage problem solution. Due to the presence of the birth rate Verhulst factor, which in effect introduces interaction between the diffusing particles, a direct comparison is not legitimate. In practice however, the Verhulst factor varies relatively little after the early stage simulation steps during which a steady total population is achieved. This explains why Fig. 4 closely resembles the first-passage survival time distribution.

Figure 5 shows the occurence frequency of the different locations of the spikes in 1600 repeated simulations. The simulation parameters are the same as those of Fig. 4 with $N_{max}=4.5\times10^4$. From the results, we can conclude that location l_s is indeed random, as our analysis would suggest. The formation of the spiked population configurations clearly reduces the diversity in lineage, as a large number of subpopulations are in effect wiped out during the spike formation. This provides a clear mechanism for the so-called Eve effect [11,12], where population evolves to have fewer and fewer common ancestors. Again the dynamics of this mechanism would closely follow that of the first-passage problem. Different power laws were observed for early and late stage simulations of the Eve effect [11,12]. This difference, according to our analysis, could be linked to the variation within (and the lack of) the Verhulst factor at the early and late stage simulations.

Finally, for cases where $T>2$, we find qualitatively similar anomalies in simulations. The analytical formulation follows a similar line as presented here, with the population $n(l_1,l)$ generalizing to $n(l_1,l_2,l)$.

IV. CONCLUSION

To conclude, we have shown by means of exact analytic solution and computer simulation that, in the asexual Penna model, a series of anomalies exist which may have affected all similar Penna simulations in the past. We have characterized these anomalies and their associated demographic distributions. Future simulations of the asexual Penna model need to pay special attention to these anomalies if reliable results are to be obtained. Our analysis also suggest that the so-called Eve effect arises in the Penna model via a mechanism similar to the first-passage problem. Similar effects in the sexual Penna model is under our ongoing investigation.

ACKNOWLEDGMENTS

The authors wish to ackowledge M. E. Cates and J. B. Coe for helpful discussions.

[1] T. J. P. Penna, J. Stat. Phys. **78**, 1629 (1995).
[2] D. Stauffer, S. Moss de Oliveira, P. M. C. de Oliveira, and J. S. Sa Martins, *Biology, Sociology, Geology by Computational Physicists* (Elsevier, Amsterdam, 2006).
[3] M. Rose, *Evolutionary Biology of Aging* (Oxford University Press, New York, 1991).
[4] P. B. Medawar, *An Unsolved Problem of Biology* (H. K. Lewis, London, 1952).
[5] B. Charlesworth, J. Theor. Biol. **210**, 47 (2001).
[6] J. B. Coe, Y. Mao, and M. E. Cates, Phys. Rev. E **70**, 021907 (2004).
[7] K. Malarz, Int. J. Mod. Phys. C **11**, 309 (2000).
[8] R. M. C. Almeida and G. L. Thomas Int. J. Mod. Phys. C, **11**, 1209 (2000).
[9] J. B. Coe and Y. Mao, Phys. Rev. E **72**, 051925 (2005).
[10] A. Laszkiewicz, S. Cebrat, and D. Stauffer, Adv. Complex Syst. **8**, 1 (2005).
[11] D. Makowiec, J. Dabkowski, and M. Groth, Physica A **273**, 169 (1999).
[12] M. Sitarz and A. Maksymowicz, Int. J. Mod. Phys. C **16**, 1917 (2005).
[13] P. M. C. de Oliveira, S. Moss de Oliveira, and D. Stauffer, Theory Biosci. **116**, 3 (1997).
[14] J. B. Coe, Y. Mao, and M. E. Cates, Phys. Rev. Lett. **89**, 288103 (2002).
[15] J. B. Coe and Y. Mao, Phys. Rev. E **67**, 061909 (2003).
[16] E. Schrödinger, Phys. Z. **16**, 289 (1915); see also S. Redner, *A Guide to First-Passage Processes* (Cambridge University Press, Cambridge, 2001).