Journal of Mathematical Biology manuscript No.
(will be inserted by the editor)

Leonardo P. Maia

The dynamical way to mutation-selection balance of an infinite population evolving on a truncated fitness landscape

Received: / Revised version: –

Abstract. This paper presents the exact analytical solution, valid for all generations and initial conditions, for the frequency distribution of haploids with infinite-sites genome carrying a given number of mutations in a population evolving deterministically on a truncated fitness landscape. This landscape is a generalization of the single sharp peak one, widely used in quasispecies theory, although here there are no reverse mutations.

1. Introduction

In general, it is very hard to find exact analytical results in population genetics models of selection. Although the behavior of a population of competing genes is usually modeled in a very simplified way, additional assumptions like assuming weak effects of mutation and/or selection are often needed in order to turn the models tractable. Even so, in many cases only stationary solutions can be found. But there are experiments where the population under study can only be monitored for periods significantly shorter than the mean life time of its organisms, like the one described in [10], page 112. The equilibrium condition is never attained in such a situation. Besides that, the idea of a population reaching an equilibrium state by evolving for a long time in a static environment may prove inadequate to describe many systems.

In this context, it is very useful to know the dynamical behavior of population genetics models. Two of the very rare works along these lines are [5] and [11], and [22] is a interesting review about dynamic fitness landscapes. We [13] recently found the solution for the full dynamics of a deterministic

Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, Caixa Postal 668, 13560-970 São Carlos SP, Brazil
e-mail: lpmaia@icmc.usp.br,lpmaia@gmail.com

The author acknowledges J. F. Fontanari for critically reading the manuscript and the Brazilian agency FAPESP for financial support. This work was developed while the author was in the Instituto de Física de São Carlos, Universidade de São Paulo, Brazil.

Key words: truncation selection – single sharp peak – Eulerian numbers – mutation-selection balance

model of organisms evolving on a multiplicative fitness landscape using a simple generating function.

This work is a new contribution along these lines. The exact evolution dynamics of an infinite population of haploid organisms on a truncated fitness landscape was determined for all generations and initial conditions. Although the general solution is a bit cumbersome, very simple expressions were found for the stationary state and for the complete dynamics of a simpler case, the single sharp peak landscape. After describing the properties and relevance of truncation selection in section 2, the details of the population genetics model and its solution are given in sections 3 and 4, respectively. Following the conclusions in section 5, the basic properties of special numbers involved in the solution, the Eulerian numbers, are described in appendix A.

2. The truncated fitness landscape

The truncated fitness landscape incorporates synergetic epistasis in a extreme way. The fitness of an individual is constant for all mutational loads with k or fewer mutations. Above this threshold, the next mutation reduces the fitness by a factor of $1 - s$, where s is a selective coefficient, but after that the fitness becomes again insensible to mutations. In symbols,

$$w_j = \begin{cases} 1, & \text{if } j \leq k \\ 1 - s, & \text{if } j > k \end{cases} . \tag{1}$$

The fitness of a sequence depends only on the total number of mutations in all its genes. So at the phenotypical level mutations in different genes cannot be distinguished. If $s = 1$, there is a "sudden death" effect: a single mutation can instantaneously kill an individual until then "healthy". It is possible that truncation selection acts on the evolutionary dynamics of repetitive sequences in eukaryotes, as discussed in [4]. Some bounds on error thresholds in landscapes slightly more general than the truncated one were found in [21]. Actually, there are very few studies of truncation selection in the sense of this paper until now, although some works can be found with the same two-class model here employed but using other genotype-phenotype mappings, as in the modeling of neutral networks.

The particular case when just one genotype has selective advantage and all others are equally misfit, $k = 0$ and $s \neq 1$, is much more well studied. There is a profusion of terminologies for this landscape in the literature, including single sharp peak, sharply-peaked, singly-peaked, isolated peak and master sequence landscape. The first one will be adopted with the hope this choice is less subject to misinterpretations. An individual with the fittest genotype is called a master sequence.

The single sharp peak landscape was originally proposed by Manfred Eigen, associated with his quasispecies theory for prebiotic evolution [7]. Since then, this landscape has been associated with the determination of error thresholds [19]. It was the model of choice in the first study of an

error threshold in finite populations [16]. An useful review on all this topics is [2]. The literature up to 1989 on the quasispecies theory, including the main applications of the single sharp peak landscape, was reviewed in [8]. However, it is worth stressing that standard quasispecies theory assumes finite genome size, in contrast with this paper.

The single sharp peak landscape was also used to describe evolution of the coliphage Qβ [6], and to study the imperfect replication of viruses [17], [14], and RNA [18]. But despite its popularity it is nowadays considered just a toy model for biological evolution and its effective usefulness is controversial [15], [3]. It should be taken seriously only as a possible model to describe the behavior of restrict regions of the genome like the ones that code the few sites in the active center of an enzyme, easily disabled by any mutation. Even so, many error thresholds calculations dangerously assume the whole genome evolves under that landscape. Besides that, in some approximations (there is no exact analytical solution for the quasispecies model even under such a simple case) it is even possible to say strong selection makes the population stable with any mutation rate [1], reinforcing severe criticisms on the error threshold concept [20], [21].

It is well known (see, e.g., [21]) that the stationary concentration of the master sequence in an infinite population of individuals with infinite genome is

$$C_0(\infty) = \frac{1 - e^U(1-s)}{s\, e^U} \tag{2}$$

and that the stationary mean fitness is $w(\infty) = e^{-U}$. Until this work, no simple expression was known for the asymptotic concentration of mutants, $C_{m>0}(\infty)$. Both results assume $C_0(0) \neq 0$ and that the mutation rate is smaller than the error threshold, $U < -\ln(1-s)$. If at least one of these conditions is not satisfied, $C_i(\infty) = 0\ \forall i$ and $w(\infty) = 1 - s$.

3. The population genetics model

The model describes the deterministic evolution of an infinite population subject to mutation and selection. Even so, it is instructive to describe the dynamics as if the population was finite and Wright-Fisher sampling could be used, what reveals a nice probabilistic interpretation.

The individuals reproduce asexually in discrete time, without superposition of generations. Each member of generation t (son) is chosen by sorting an individual of generation $t-1$ (father) with probability proportional to its fitness. The son inherites all mutations of its father and gets an additional number k of them (interpreted as replication failures) sampled from a Poisson distribution with mean U,

$$M_k = e^{-U}\frac{U^k}{k!}. \tag{3}$$

The individuals are haploid. Each one is represented by a single sequence of infinite genes (Kimura's infinite sites model) and therefore no gene can ever mutate more than one time and there are no reverse mutations.

If $C_i(t)$ is the fraction of sequences that carry i mutations at time t, each one with fitness w_i, the probability that a sequence carrying j mutations be chosen to reproduce is

$$P_j(t) = \frac{C_j(t)w_j}{w(t)}, \tag{4}$$

where

$$w(t) = \sum_{k=0}^{\infty} C_k(t)w_k \tag{5}$$

is the mean fitness at time t. The phenotypical evolution of the population is described by the convolution equation

$$C_i(t) = \sum_{j=0}^{i} P_j(t-1)M_{i-j} = \sum_{j=0}^{i} \frac{C_j(t-1)w_j}{w(t-1)}\, e^{-U} \frac{U^{i-j}}{(i-j)!}, \tag{6}$$

used for the first time by Kimura and Maruyama [12].

In principle, the r.h.s. of Eq. (6) can be useful in studies of finite populations too since it is the conditional probability of an individual with j mutations in generation $t-1$ be chosen to reproduce and generate an individual with i mutations in generation t, given the concentrations at $t-1$. Only in the infinite population limit this probability is exactly equal to $C_i(t)$ (absence of genetic drift).

4. The solution

The truncated fitness landscape allows the mean fitness of the population to be expressed in terms of a generating function, what turns the problem solvable just like in [13]. The generating function here employed is defined as

$$G(z,t) = \sum_{i=0}^{\infty} C_i(t)z^i. \tag{7}$$

Given it, the frequency of individuals carrying i mutations at instant t can be easily found as the coefficient of z^i in the Taylor's power series expansion of $G(z,t)$,

$$C_i(t) = \frac{1}{i!}\left\{\frac{\partial^i}{\partial z^i}G(z,t)\right\}_{z=0}. \tag{8}$$

A recurrence equation for the generating function is found by multiplying both sides of Eq. (6) by z^i and summing i from 0 up to ∞ so that

$$G(z,t) = \frac{e^{zU}\left[s\sum_{j=0}^{k} z^j C_j(t-1) + (1-s)G(z,t-1)\right]}{e^{U}\left[s\sum_{j=0}^{k} C_j(t-1) + (1-s)\right]}. \tag{9}$$

Recursively back to $t = 0$,

$$G(z,t) = \left\{ s\, e^{zU} \sum_{j=0}^{k}\sum_{i=0}^{t-1}\sum_{l=0}^{k-j} z^j C_j(0)[e^{zU}(1-s)]^{t-1-i}\frac{(iUz)^l}{l!} \right.$$
$$\left. +[e^{zU}(1-s)]^t G(z,0) \right\}$$
$$\div \left\{ s\, e^{U} \sum_{j=0}^{k}\sum_{i=0}^{t-1}\sum_{l=0}^{k-j} C_j(0)[e^{U}(1-s)]^{t-1-i}\frac{(iU)^l}{l!} + [e^{U}(1-s)]^t \right\}. \tag{10}$$

The distribution of frequencies comes from Eq. (8),

$$C_m(t) = \left\{ s \sum_{j=0}^{\min(k,m)}\sum_{i=0}^{t-1}\sum_{l=0}^{\min(k,m)-j} C_j(0)\frac{U^{m-j}}{(m-j)!}(1-s)^{t-1-i} \right.$$
$$\left. \times \binom{m-j}{l} i^l (t-i)^{m-j-l} + (1-s)^t \sum_{j=0}^{m} C_j(0)\frac{(Ut)^{m-j}}{(m-j)!} \right\}$$
$$\div \left\{ s\, e^{U} \sum_{j=0}^{k}\sum_{i=0}^{t-1}\sum_{l=0}^{k-j} C_j(0)[e^{U}(1-s)]^{t-1-i}\frac{(iU)^l}{l!} + [e^{U}(1-s)]^t \right\}. \tag{11}$$

The concentrations of the fittest individuals are somewhat simpler,

$$C_{m\le k}(t) = \left\{ \sum_{j=0}^{m} C_j(0)\frac{(Ut)^{m-j}}{(m-j)!} \right\}$$
$$\div \left\{ s\, e^{U} \sum_{j=0}^{k}\sum_{i=0}^{t-1}\sum_{l=0}^{k-j} C_j(0)[e^{U}(1-s)]^{t-1-i}\frac{(iU)^l}{l!} + [e^{U}(1-s)]^t \right\}. \tag{12}$$

The mean fitness of the population is

$$w(t) = 1 - s + s\sum_{m=0}^{k} C_m(t) = 1 - s + \left\{ s\sum_{j=0}^{k} C_j(0)\sum_{l=0}^{k-j}\frac{(Ut)^l}{l!} \right\}$$
$$\div \left\{ s\, e^{U} \sum_{j=0}^{k}\sum_{i=0}^{t-1}\sum_{l=0}^{k-j} C_j(0)[e^{U}(1-s)]^{t-1-i}\frac{(iU)^l}{l!} + [e^{U}(1-s)]^t \right\} \tag{13}$$

and at any time the last two equations depend only on the initial concentrations of the fittest sequences. It is not possible to simplify these expressions in the general case. But the $k = 0$ and $t \to \infty$ cases allow analytical progress.

4.1. The single sharp peak case

From Eq. (12), the master sequence concentration in the single sharp peak landscape is

$$C_0^*(t) = \frac{[1 - e^U(1-s)]C_0(0)}{s\,e^U C_0(0) + \{1 - e^U + s\,e^U[1 - C_0(0)]\}\{e^U(1-s)\}^t} \tag{14}$$

and from (13) the mean fitness is given either by

$$w^*(t) = 1 - s + sC_0^*(t), \tag{15}$$

or, more explicitly, by

$$w^*(t) = \frac{sC_0(0) + (1-s)\left\{1 - e^U + s\,e^U[1 - C_0(0)]\right\}\left\{e^U(1-s)\right\}^t}{s\,e^U C_0(0) + \{1 - e^U + s\,e^U[1 - C_0(0)]\}\{e^U(1-s)\}^t}. \tag{16}$$

Both the error threshold and the stationary solution (2) follow by a straighforward analysis of the asymptotic behavior of Eqs. (14) and (15).

The frequencies of the mutants are found from Eq. (11),

$$C_{m>0}(t) = \left\{1 - e^U(1-s)\right\} \times \left\{ \frac{s}{1-s} C_0(0) \frac{U^m}{m!} \sum_{j=1}^{t} j^m (1-s)^j + (1-s)^t \sum_{j=0}^{m} C_j(0) \frac{(Ut)^{m-j}}{(m-j)!} \right\} \div \left\{ s\,e^U C_0(0) + \left\{1 - e^U + s\,e^U[1 - C_0(0)]\right\}\left\{e^U(1-s)\right\}^t \right\}. \tag{17}$$

The first sum in this equation can be formally expressed in terms of generalized Eulerian numbers (see the second reference in [9]), but it is not useful due to the complexity of the procedure. Using Eq. (23) from appendix A, the stationary concentration of the mutants below the error threshold is easily found to be

$$C_{m>0}(\infty) = \left\{ \frac{1 - e^U(1-s)}{s\ e^U} \right\} \left\{ \frac{(U/s)^m}{m!} \right\} \mathcal{P}_m(1-s), \tag{18}$$

where

$$\mathcal{P}_m(x) \equiv \sum_{k=0}^{m-1} E_{m,k} x^k \tag{19}$$

and the Eulerian numbers $E_{m,k}$ are discussed in appendix A.

4.2. The general stationary state solution

With no loss of generality, one can assume $C_0(0) \neq 0$ (otherwise k is redefined in terms of the first non-null concentration). It is not hard to see from (12) that for $0 \leq m < n < k$

$$\lim_{t\to\infty} \frac{C_m(t)}{C_n(t)} = 0. \tag{20}$$

So, asymptotically, all sequences with less than k mutations disappear whatever be the mutation rate and the selection coefficient. Consequently, the stationary state of the population on a truncation selection landscape is the same one found for the single sharp peak landscape with the individuals carrying k mutations playing the role of master sequences. Of course, the error threshold is the same too, as could be guessed by looking at the general solution. Obviously, explicit calculations (not shown) confirm this observations. So, if $e^U(1-s) < 1$ and $\mathcal{P}_0(x) \equiv 1$, $C_{m<k}(\infty) = 0$ and

$$C_{m\geq k}(\infty) = \left\{\frac{1-e^U(1-s)}{s\ e^U}\right\}\left\{\frac{(U/s)^{m-k}}{(m-k)!}\right\}\mathcal{P}_{m-k}(1-s). \tag{21}$$

This result is not surprising because the truncated landscape is flat below k and it is well known that a flat landscape cannot prevent the probability mass from spreading. Only the entropic barrier at k can possibly do that, if the error threshold is not crossed.

5. Conclusions

The dynamical behavior of an infinite population of asexual haploid organisms evolving on a truncated fitness landscape (a generalization of the most popular toy landscape, the single sharp peak one) was found for any initial condition. In particular, the dynamics of the single sharp peak case is described by simple expressions and its stationary state also describes the mutation-selection balance reached under truncation selection. Far from the error threshold condition, convergence to the stationary concentrations can be very fast due to exponential terms in time dependence. In such conditions, the already known Eq. (2) and the mean fitness result below it may be used to estimate the parameters of the model. However, there is evidence that some populations actually evolve close to their error thresholds, where the full dynamical solution can be useful.

Even if the fittest sequences cannot be distinguished one from another, the first equality in (13) shows that it is possible to get information about the selection parameter at any time if the mean fitness of the population and the total concentration of fittest organisms are known. The mutation rate can be found by looking at the state of the system in different times, independently of the initial conditions.

Finally, it is worth noting that, under strong selection ($s = 1$), a population evolving on the truncated landscape here studied reaches the same

stationary state (the concentration follows a Poisson distribution with parameter U) reached under a multiplicative landscape [13], as long as asymptotically there is only one kind of sequence favored in both models. So these two landscapes are indistinguishable in this situation.

A. Appendix - On Eulerian numbers

The Eulerian numbers $E_{n,k}$ are special numbers specified by two integers, like binomial coefficients. And although they are much less famous than their binomial cousins, Eulerian numbers also generate a symmetric triangle like Pascal's one and have an interesting combinatorial interpretation: $E_{n,k}$ is the number of permutations $\pi_1\pi_2...\pi_n$ of $\{1, 2, ..., n\}$ that have k places where $\pi_j < \pi_{j+1}$ ("rises"). So k can assume any value from 0 to $n-1$ and $\sum_{k=0}^{n-1} E_{n,k} = n!$. The symmetry $E_{n,k} = E_{n,n-1-k}$ is due to the fact that the permutation $\pi_1\pi_2...\pi_n$ can have $n-1-k$ "rises" if and only if its reflection $\pi_n...\pi_2\pi_1$ has k "rises".

By alternately differentiating and multiplying by x the usual geometric series for $\mid x \mid < 1$

$$\frac{1}{1-x} = \sum_{i=0}^{\infty} x^i \tag{22}$$

one gets

$$\frac{x}{(1-x)^{n+1}} \sum_{k=0}^{n-1} E_{n,k} x^k = \sum_{i=1}^{\infty} i^n x^i. \tag{23}$$

This procedure reveals that Eulerian numbers can be generated with the recurrence

$$E_{n,k} = (k+1)E_{n-1,k} + (n-k)E_{n-1,k-1}. \tag{24}$$

The first Eulerian numbers are shown in table 1.

References

1. Alves, D., Fontanari, J. F.: A population genetics approach to the quasispecies model. Phys. Rev. E **54** (1996) 4048-4053. Available at http://arxiv.org/abs/cond-mat/9605160
2. Baake, E., Gabriel, W.: Biological evolution through mutation, selection, and drift: an introductory review. Ann. Rev. Comp. Phys. **9** 203-264 (D. Stauffer ed., World Scientific 2000). Available at http://arxiv.org/abs/cond-mat/9907372
3. Charlesworth, B.: Mutation-selection balance and the evolutionary advantage of sex and recombination. Genet. Res. Camb. **55** (1990) 199-221
4. Charlesworth, B., Sniegowski, P., Stephan, W.: The evolutionary dynamics of repetitive DNA in eukaryotes. Nature **371** (1994) 215-220
5. Dawson, K. J.: The dynamics of infinitesimally rare alleles, applied to the evolution of mutation rates and the expression of deleterious mutations. Theor. Pop. Biol. **55** (1999) 1-22

6. Domingo, E., Sabo, D., Taniguchi, T., Weissmann, C.: Nucleotide sequence heterogeneity of an RNA phage population. Cell **13** (1978) 735-744
7. Eigen, M.: Selforganization of matter and the evolution of biological macromolecules. Naturwiss. **58** (1971) 465-523
8. Eigen, M., McCaskill, J., Schuster, P.: Molecular quasi-species. J. Phys. Chem. **92** (1988) 6881-6891; The molecular quasi-species. Adv. Chem. Phys. **75** (1989) 149-263
9. Graham, R. L., Knuth, D. E., Patashnik, O.: Concrete Mathematics - A Foundation for Computer Science (Addison-Wesley, 1994), Second Edition; http://www.mathpages.com/home/kmath464.htm
10. Hartl, D. L., Clark, A. G.: Principles of Population Genetics (Sinauer Associates Inc., Sunderland 1989), Second Edition
11. Johnson, T.: The approach to mutation-selection balance in an infinite asexual population, and the evolution of mutation rates. Proc. R. Soc. Lond. B **266** (1999) 2389-2397; Theoretical Studies of the Interaction Between Deleterious and Beneficial Mutations (PhD thesis, University of Edinburgh 2000). Available at http://homepages.ed.ac.uk/tobyj
12. Kimura, M., Maruyama, T.: The mutational load with epistatic gene interactions in fitness. Genetics **54** (1966) 1337-1351
13. Maia, L. P., Botelho, D. F., Fontanari, J. F.: Analytical solution of the evolution dynamics on a multiplicative-fitness landscape. J. Math. Biol. **47** (2003) 453-456
14. Martinez-Salas, E., Ortin, J., Domingo, E.: Sequence of the viral replicase gene from foot and mouth disease virus C_1-Santa Pau (C-S8). Gene **35** (1985) 55-61
15. Maynard Smith, J.: Models of evolution. Proc. R. Soc. Lond. B **219** (1983) 315-325
16. Nowak, M., Schuster, P.: Error thresholds of replication in finite populations: mutation frequencies and the onset of Muller's ratchet. J. Theor. Biol. **137** (1989) 375-395
17. Ortin, J., Nájera, R., Lopez, C., Davila, M., Domingo, E.: Genetic variability of Hong Kong (H3N2) influenza viruses: spontaneous mutations and their location in the viral genome. Gene **11** (1980) 319-331
18. Spiegelman, S., Haruna, I., Holland, I. B., Beaudreau, G., Mills, D. R.: The synthesis of a self-propagating and infectious nucleic acid with a purified enzyme. Proc. Natl. Acad. Sci. USA **54** (1965) 919-927
19. Swetina, J., Schuster, P.: Self-replication with errors: a model for polynucleotide replication. Biophys. Chem. **16** (1982) 329-345
20. Wagner, G. P., Krall, P.: What is the difference between models of error thresholds and Muller's ratchet? J. Math. Biol. **32** (1993) 33-44
21. Wiehe, T.: Model dependency of error thresholds: the role of fitness functions and contrasts between the finite and infinite sites models. Genet. Res. Camb. **69** (1997) 127-136
22. Wilke, C. O., Ronnewinkel, C., Martinetz, T.: Dynamic fitness landscapes in molecular evolution. Phys. Rep. **349** (2001) 395-446. Available at http://arxiv.org/abs/physics/9912012

Table 1. Euler's triangle

n	$E_{n,0}$	$E_{n,1}$	$E_{n,2}$	$E_{n,3}$	$E_{n,4}$	$E_{n,5}$	$E_{n,6}$	$E_{n,7}$	$E_{n,8}$
0	1								
1	1	0							
2	1	1	0						
3	1	4	1	0					
4	1	11	11	1	0				
5	1	26	66	26	1	0			
6	1	57	302	302	57	1	0		
7	1	120	1191	2416	1191	120	1	0	
8	1	247	4293	15619	15619	4293	247	1	0