Collapsing the State Space

Applying Markov Analysis to Evolutionary Systems

Lionel Barnett

Centre for Computational Neuroscience and Robotics
Centre for the Study of Evolution
Department of Cognitive and Computing Sciences
University of Sussex, Brighton BN1 9QH, UK
lionelb@cogs.susx.ac.uk

Abstract

Evolutionary systems may often be accurately modelled by Markov processes. However the state space invariably turns out to be vast and multi-dimensional, thus limiting the application of Markov theory to broad abstraction rather than specific problems. One notable exception is the analysis of error thresholds in finite populations by (Nowak & Schuster 1989), where a (seemingly unjustifiable) approximation is made to "collapse" the state space, reducing the problem to an analytically tractable form. In this paper we outline Nowak and Schuster's analysis and discuss the methodology of their approach.

Error thresholds for Finite Populations

(Nowak & Schuster 1989) investigated the extension of established results from "quasispecies" theory (Eigen *et. al.* 1989) on error thresholds for infinite populations to finite populations. The basic problem is as follows: we are given a "single spike" fitness landscape of binary genotypes of sequence length ν. All genotypes have fitness 1 except for the genotype consisting of all zero's (the *master* genotype or *optimum*), which has fitness $\sigma > 1$. Genotypes Hamming distance α from the optimum are said to belong to the *error class* Γ_α - the Γ_α for $\alpha > 1$ constitute the *error tail*.

Consider a fixed-size population of N genotypes evolving via fitness-proportional selection[1] and mutation at a per-locus rate of μ ($0 \le \mu \le ½$). There is no recombination. The observed long-term behaviour of such a system is as follows: at low mutation rates the population clusters around the optimum (Fig 1a). At higher mutation rates more genotypes are to be found at a small (Hamming) distance from the optimum (Fig 1b). Beyond a critical mutation rate, the *error threshold,* the population "loses" the optimum altogether and drifts randomly around the landscape[2] (Fig 1c). In the infinite population limit the error threshold may be calculated from quasispecies theory using perturbation methods (Eigen *et. al.* 1989). For finite populations the error threshold is less easy to define, let alone calculate. Nevertheless, there is still a sharp transition between long-term behaviours in the sense that the transition (for reasonably long sequence length ν) occurs within a very small range of mutation rates.

To analyse the transition we must examine the distribution (over time) of the number of optimum genotypes, $\pi_i \equiv \mathbf{P}(\hat{X} = i)$, where the random variable $\hat{X}$ represents the number of optimum genotypes in the long term. (Nowak & Schuster 1989) found that at low mutation rates the distribution peaks at some characteristic value of i (Fig 2a). At high mutation rates the distribution decreases monotonically from i = 0 (Fig 2c). At intermediate mutation rates the distribution develops a second peak at i = 0 (Fig 2b). The authors then *define* the error threshold to be that value of μ at which the distribution changes from monotone decreasing to one with a peak at i > 0.

Now it is apparent that we could calculate the distribution π_i if it were true that the random variables X(t) representing the number of optimum genotypes at (discrete or continuous) time t constituted a Markov process. The distribution π_i would then be simply the stationary distribution of the process (Karlin & Taylor 1975). However it is clear that the Markov property does not hold, for the following reason: while the probability that a genotype be selected for replication depends only on whether it is of the optimum type or in the error tail, the probability that a genotype in the error tail "back-mutates" to the optimum type depends, in

[1] The exact selection algorithm in effect does not alter the qualitative phenomena; thus selection may be roulette-wheel, tournament, etc. as long as the *expected* number of offspring of a genotype is proportional to its fitness. The algorithm may, in addition, be discrete or continuous time.

[2] This implies that the distribution of number of genotypes among the error classes is *binomial*, as the α'th error class occupies a fraction $2^{-\nu}\binom{\nu}{\alpha}$ of the landscape; cf. Fig 1c.

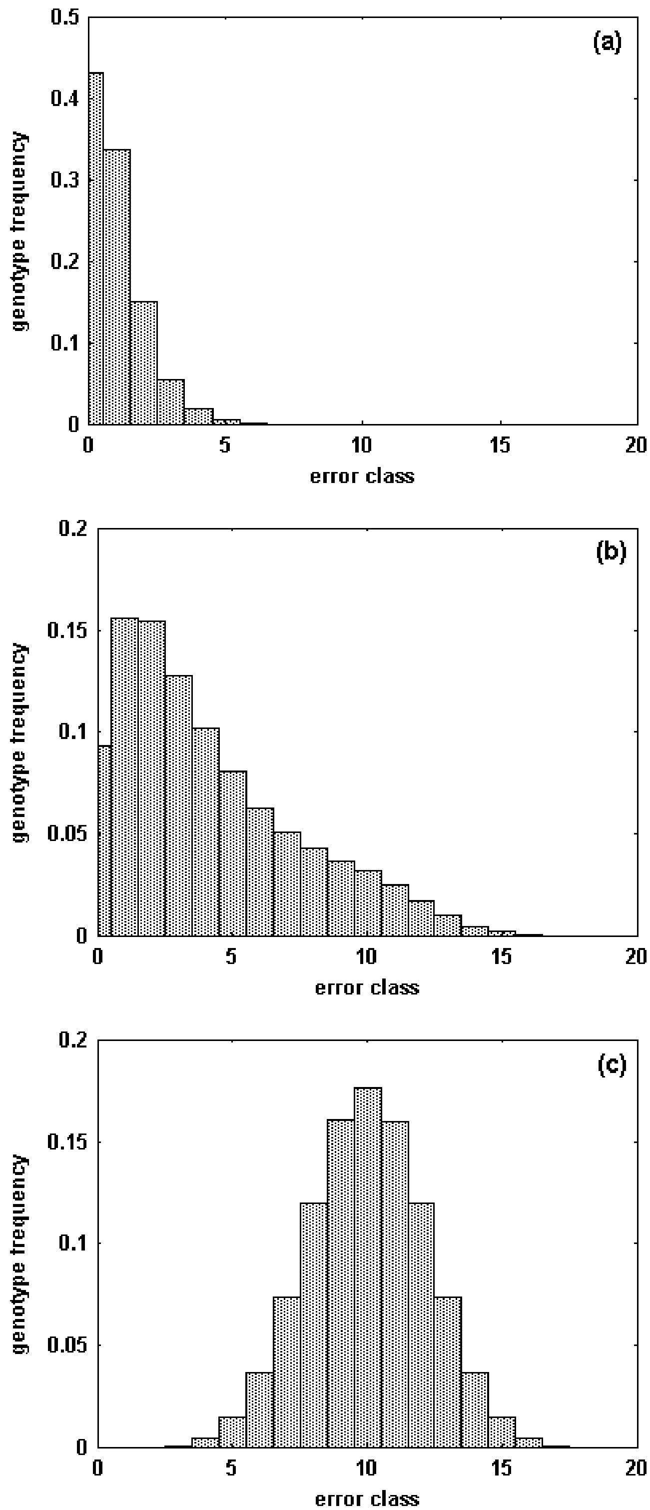

Fig 1. *Distribution of genotype frequency over error classes for $\sigma = 5$, $\nu = 20$, $N = 200$ and per-locus mutation rates (a) $\mu = 0.03$, (b) $\mu = 0.061$ and (c) $\mu = 0.07$. Simulation was over 100,000 generations of a roulette-wheel fitness-proportional selection algorithm.*

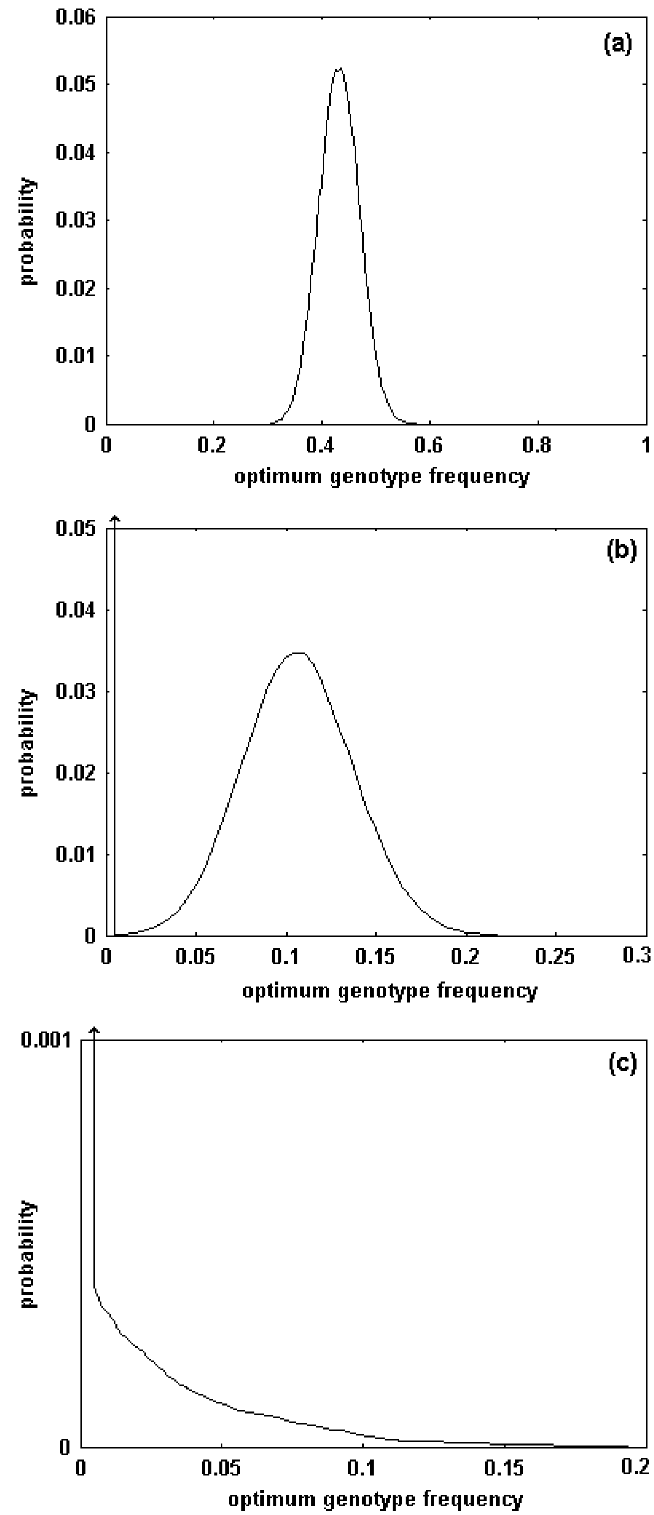

Fig 2. *Stationary probability distribution π_i of optimum (error class 0) genotype frequency for the same simulations as in Fig 1. The error threshold is approximately $\mu = 0.063$.*

addition, on how many 1's it has; i.e. its error class. Nowak and Schuster address this issue by making what on the face of it is an unjustifiable assumption: that the distribution of genotypes is *uniform* in the error tail. Under this assumption both selection and mutation probabilities depend only on the number of optimum genotypes and the Markov property holds. The authors then carry through the Markov analysis for a particular evolutionary algorithm [a continuous birth-death model from population genetics - see (Moran 1958)] for which the stationary distribution is explicitly solvable, thus arriving at what turns out to be a very accurate estimate for the error threshold.

So why should this scheme work at all? It is clear from Figs. 1a and 1b that the assumption of a uniform distribution of genotypes in the error tail is manifestly false. A point apparently missed by the authors, though, is that at *high* mutation rates, when the optimum is "lost", the uniform distribution assumption is actually quite sound, as is evidenced by Fig. 1c. This may explain the accuracy of their result to some degree; as long as the error threshold is approached "from above" the assumption holds good. Another point worth noting is that for reasonably long sequence length ν the probability of back-mutation from the error tail to the optimum becomes small (of the order of $2^{-\nu}$), even for genotypes a small Hamming distance from the optimum. Then ignoring back mutation entirely is a reasonable approximation and the Markov property holds without any further assumptions.

Another possible approach might be as follows: it is possible to calculate (numerically at least) the distribution of genotypes in the infinite population limit (van Nimwegen *et. al.* 1997). This limiting distribution is quite a good approximation to the finite population case for reasonably large populations, although it is not clear what "reasonably" large might mean. Thus, rather than assuming a uniform distribution of genotypes in the error tail we could assume instead the infinite population limit. Preliminary tests by this author suggest that this can give a significantly more accurate approximation to the stationary distribution of the optimum than the cruder uniform distribution assumption, particularly near the error threshold .

Methodological Issues

It is worth examining how the procedure outlined above tackles the issue of state space size and dimension. The full state space for the problem of a fixed-size population of size N evolving on a fitness landscape is the set of all possible populations. A population is naturally identified with an integer vector $\mathbf{n} = (n_g)$ indexed by all possible genotypes g, where n_g represents the number of copies of genotype g in the population. The n_g must satisfy $n_g \geq 0 \ \forall g$ and $\sum_g n_g = N$. The state space is thus vast and multi-dimensional; if sequence length is ν then the cardinality of the state space is $\binom{\nu + N}{\nu}$ which is of the order of N^{ν} for $N >> \nu$. The crucial point is that if all we are interested in is the error threshold, the only quantity we need to know is the stationary probability distribution of the frequency of optimum genotypes. Now since the single-spike landscape is "isotropic" with respect to the optimum genotype we can immediately "collapse" the state space into the frequencies of genotypes in the error classes without losing either the Markov property or the quantity we wish to measure. This is possible because mutation and selection probabilities (and hence the transition probabilities of the Markov process) depend only on error class. Thus our state space may be immediately reduced to the set of vectors $\mathbf{n} = (n_\alpha)$ indexed now by the error classes α. [If recombination were present this would no longer be true - see below.] Note that thus lose all information as to the distribution of genotypes *within* error classes - but we do not need this information for the problem at hand! The state space is then reduced still further by (cautious) approximation to an analytically tractable 1-dimensional space.

Another point that may be overlooked in the quest for quantitative results is that even if various approximations introduce quantitative inaccuracies (as they do to some degree in Nowak and Schuster's analysis), the *qualitative* picture may still hold up. Thus valuable insights may be gained into the dynamical behaviour of an evolutionary system by the introduction of simplifying assumptions; this is certainly the case for Nowak and Schuster's analysis of error thresholds.

As a further case in point this author [in preparation] has extended Nowak and Schuster's analysis to include recombination, revealing a rich and often surprising range of dynamics. This necessitated the introduction of further (quantitatively unjustifiable) assumptions, specifically because the Markov property does not hold even for recombination *within* error classes. Comparing analytical results with simulations, however, reveals that the approximation retains almost all qualitative features of interest.

Finally, it would seem to be feasible to extend these principles to the analysis of evolution on more complex landscapes, particularly if they feature analogues of the error classes. A comparable approach (although not for the purposes of Markov analysis) can be found in (van Nimwegen *et. al.* 1997).

Thus we might define a partition $\{\Gamma_\alpha | \alpha \in A\}$ of a fitness landscape with fitness function f(g) and (stochastic) mutation operator M_μ to be a ***Markov Partition*** if it satisfies:

$\forall \alpha \in A, \ \forall g, g' \in \Gamma_\alpha \forall, \ \beta \in A$ and $\forall \mu \ (0 \le \mu \le 0.5)$:

MP1 $$f(g') = f(g)$$

MP2 $$\mathbf{P}(M_\mu(g') \in \Gamma_\beta) = \mathbf{P}(M_\mu(g) \in \Gamma_\beta)$$

The quantities:

$$f_\alpha \equiv f(g) \text{ for some } g \in \Gamma_\alpha$$

and:

$$M_{\alpha\beta}(\mu) \equiv \mathbf{P}(M_\mu(g) \in \Gamma_\alpha \mid g \in \Gamma_\beta)$$

are then well-defined. An evolutionary process may then be considered as a Markov process on the state space of all integer vectors $\mathbf{n} = (n_\alpha)$ with $n_\alpha \ge 0 \ \forall \alpha$ (and $\sum_\alpha n_\alpha = N$ for fixed population size N), the transition probabilities being determined (for the particular evolutionary algorithm employed) by the f_α and $M_{\alpha\beta}(\mu)$.

The *coarsest* such partition yields the smallest and most manageable state space. If possible, depending on specific features of the fitness landscape, a coarser partition might be found by relaxing conditions **MP1** and/or **MP2** to hold approximately. A further condition to cover recombination could also be defined, although it seems doubtful that useful partitions could be found which would respect such a condition exactly.

MP1 corresponds to the statement that the Γ_α are *neutral subsets* (Barnett 1998) of the fitness landscape. A particular case of interest is where they constitute the *neutral networks* (Forst *et. al.* 1995, Huynen *et. al.* 1996, Barnett 1989) of the landscape i.e. maximal connected neutral subsets. In particular, when the neutral networks "percolate" the landscape (Forst *et. al.* 1995) it seems reasonable that **MP2** might be expected to hold to some approximation.

Acknowledgements

This paper arose out of discussions of the Neutral Networks group at the Sussex University CCNR. The author would like to thank all those who took part and Inman Harvey in particular.

References

Barnett, L. 1998. *Ruggedness and Neutrality - the NKp Family of Fitness Landscapes*. Proc. 6th Intl. Conf. on Artificial Life (in print).

Eigen, M., McCaskill, J. & Schuster, P. 1989. *The Molecular Quasispecies.* Adv. Chem. Phys. 75:149-263.

Forst, C.V., Reidys, C. & Weber, J. 1995. *Neutral Networks as Model-Landscapes for RNA Secondary-Structure Folding-Landscapes.* Lecture Notes in Artificial Intelligence, vol. 929: Advances in Artificial Life (Morán, F., Moreno, A., Merelo, J.J. & Chacón eds.), Springer-Verlag, Berlin.

Huynen, M.A., Stadler, P.F. & Fontana, W. 1996. *Smoothness Within Ruggedness: The Role of Neutrality in Adaptation*. Proc. Natl. Acad. Sci. (USA) 93:397-401.

Karlin, S. & Taylor, H.M. 1975 *A First Course in Stochastic Processes (2nd ed.).* Academic Press, New York.

Moran, P.A.P. 1958 *The Effect of Selection in a Haploid Genetic population.* Proc. Camb. Phil. Soc. 54: 463-465.

van Nimwegen, E., Crutchfield, J.P. & Mitchell, M. 1997. *Statistical Dynamics of the Royal Road Genetic Algorithm.* Santa Fe Institute Pre-print 97-04-035, Santa Fe, NM, USA.

Nowak, M. & Schuster, P. 1989. *Error Thresholds of Replication in Finite Populations - Mutation Frequencies and the Onset of Muller's Ratchet*. J. Theor. Biol. 137:375-395.