DIFFUSION OF A POPULATION OF INTERACTING REPLICATORS IN SEQUENCE SPACE

BÄRBEL M. R. STADLER

Max Planck Institute for Mathematics in the Sciences
*Inselstraße 22-26, D-04103 Leipzig, Germany**
Institut für Theoretische Chemie und Strukturbiologie, Universität Wien
Währingerstraße 17, A-1090 Wien, Austria

stadler@mis.mpg.de

We consider a simple model for catalyzed replication. Computer simulations show that a finite population moves in sequence space by diffusion analgous to the behavior of a quasispecies on a flat fitness landscape. The diffusion constant depends linearly on the per position mutation rate and the ratio of sequence length and population size.

Keywords: Replicator Dynamics, Diffusion in Sequence Space, Finite haploid populations

1. Introduction

The dynamics of finite haploid populations can be described by Eigen's quasispecies model [3]. The underlying replication mechanism is

$$\mathbb{I}_k \rightarrow 2\mathbb{I}_k \tag{1}$$

with mutation. In this model, diffusion can be observed in flat fitness landscapes [1] as well as in fitness landscapes that correspond to neutral nets [8,5]. The corresponding diffusion constant is proportional to the per position mutation rate p.

Much less is known about an analogous model for second-order replicator equations, i.e., in the case of catalyzed replication. Despite the fact that there exist quite a few computer simulations [5,4,9], the question of how and under which circumstances the population diffuses, has not been answered so far.

Serva and Peliti [12] as well as Higgs and Derrida [7] studied the limiting case of infinitely long sequences and therefore vanishing per digit mutation rate. In the context of molecular evolution, however, one is interested in the case of small populations and small chain lengths. We show here that the behavior is qualitatively different. In particular we find a diffusive motion of the sequences which is analogous to the behavior of a quasispecies population.

*Address for correspondence. Tel.: ++49 341 9959 536, Fax: ++49 341 9959 658

2. The Model

We consider a stochastic version of a second order replicator equation [11] with mutation, i.e., a replication mechanism of the form

$$\mathbb{I}_k + \mathbb{I}_j \longrightarrow \mathbb{I}_l + \mathbb{I}_k + \mathbb{I}_j \tag{2}$$

with a replication rate A_{kj}. The probability to obtain an offspring of type $\mathbb{I}_l$ from a parent of type $\mathbb{I}_k$ is Q_{lk}. The deterministic version leads to the selection-mutation equation

$$\dot{x}_k = x_k \left(\sum_j A_{kj} x_j - \sum_{i,j} A_{ij} x_i x_j \right) + \sum_{l,j} \left(Q_{kl} A_{lj} x_j x_l - Q_{lk} A_{kj} x_k x_j \right) \tag{3}$$

for the relative concentration of species $\mathbb{I}_k$ [13].

We simulate a population of N sequences of length n that are composed from an alphabet $\mathcal{A}$ consisting of $\alpha = |\mathcal{A}|$ letters. These sequences replicate according to the mechanism (2). As in Eigen's quasispecies model [2] we model mutation as independent event at each sequence position, i.e.,

$$Q_{kl} = (1-p)^{n-d(k,l)} \left(\frac{p}{\alpha - 1} \right)^{d(k,l)} \tag{4}$$

where $d(k,l)$ denotes the Hamming distance of $\mathbb{I}_k$ and $\mathbb{I}_l$, and p is the single digit mutation frequency, also known as *error rate*. The replication rate A_{kj} depends on the mutual relationships of the two sequences $\mathbb{I}_k$ and $\mathbb{I}_j$. For simplicity we assume that A_{kj}, like Q_{kl}, depends only on the Hamming distance: $A_{kj} = 1 - d(k,j)/n$.

In order to save computer resources we use an approximate simulation scheme instead of an exact simulation, e.g. using the Gillespie algorithm [6]: Two out of the N sequences are chosen randomly from a tank reactor at each time step. The first sequence $\mathbb{I}_k$ acts as a template and gets replicated with probability A_{kj}. The second sequence $\mathbb{I}_j$ acts as catalyst, i.e., it determines the rate constant. Then each sequence position is mutated independently with probability p. After each successful replication event a randomly chosen sequence is removed from the tank in order to keep the number of sequences constant.

3. Simulations

Let $\mathbb{P}(0) = \{\mathbb{P}^k(t) | k = 1, \ldots, N\}$ be the population of strings at timestep t. We write $\mathbb{P}_i^k$ for the ith sequence position of string k in the population. Initially, the tank reactor is initialized with a random population $\mathbb{P}(0)$.

The *diversity* of the population $\mathbb{P}$ is defined as the average pairwise difference of the sequences in the tank reactor:

$$\delta = \frac{1}{n\binom{N}{2}} \sum_{k<l} d(\mathbb{P}^k, \mathbb{P}^l) = \frac{1}{n\binom{N}{2}} \sum_{k<l} \sum_{i=1}^{n} \left(1 - \delta(\mathbb{P}_i^k, \mathbb{P}_i^l)\right) \tag{5}$$

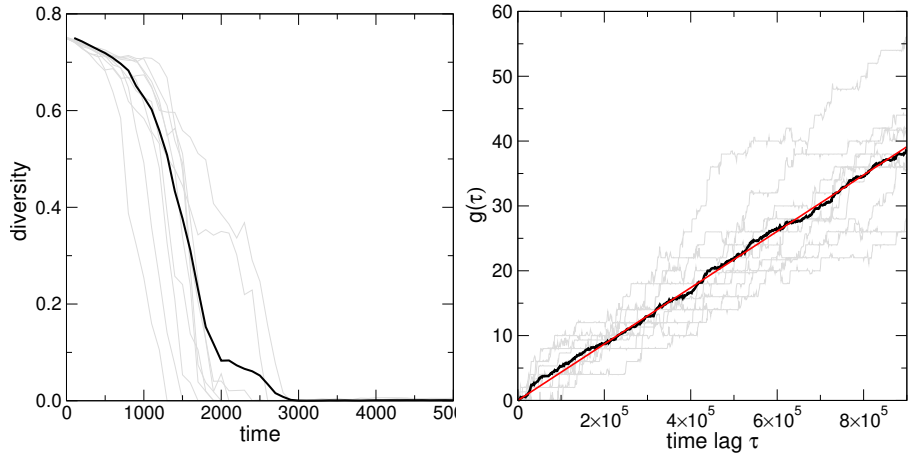


Fig. 1. Time-development of the diversity (l.h.s.) and displacement of the profile $g(\tau)$, equ.(8), on the r.h.s. for 10 independent simulations with $N = 80$ individuals, $n = 80$, $p = 0.032$. Individual simulations are shown in grey, the mean value as black line. The diffusion constant D is the slope of the averaged $g(\tau)$ curve. A transient period of 10^5 steps was removed for the computation of $g(\tau)$.

At $t = 0$ the diversity equals the average distance between two random sequences, i.e., $\delta = 1 - \frac{1}{\alpha}$. After a transient period the diversity sharply drops to almost zero when the population collapses to a mutant cloud surrounding a single "master" sequence, Fig. 1.

The papers [12,7] consider the limit of infinite sequence length $n \to \infty$ and large populations N. Therefore, they describe only the behavior before the population collapses around a single sequence, the waiting time for which event diverges with population size.

The *tank profile* $\mathbf{p}$ is the $\alpha \times n$ vector that lists the frequency of each letter at each sequence position:

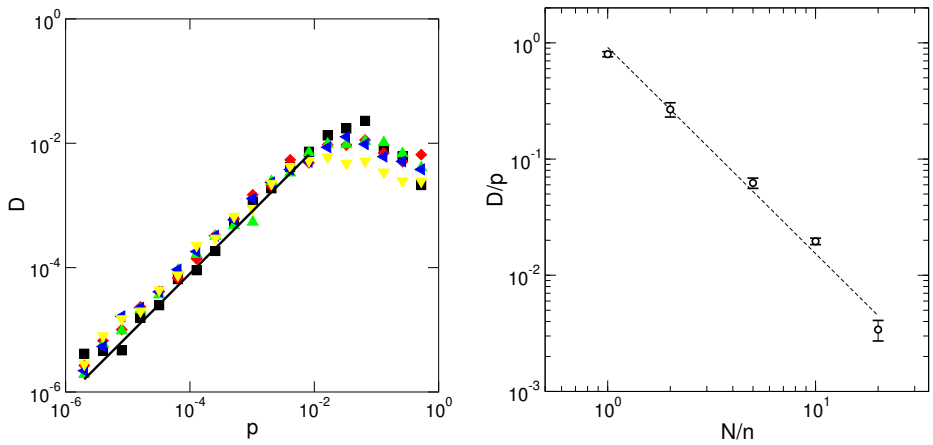
$$\mathbf{p}_{\alpha(i-1)+(j-1)}(t) = \frac{1}{N}\sum_{k=1}^{N}\delta(\mathbb{P}_i^k(t), \mathsf{a_j}) \tag{6}$$

It describes the composition of the population separately for each position. This is justified since the contribution of the individual sequence positions to the Hamming distance, and hence to Q_{kl} and A_{kl}, is independent.

The *diffusion constant* of the population at timestep t is then defined as

$$D = \lim_{\tau\to 0}\frac{\|\mathbf{p}(t+\tau) - \mathbf{p}(t)\|^2}{\tau} \tag{7}$$

where $\|\,.\,\|$ denotes the Euclidean norm. In practice, of course, D is determined as



(a) (b)

Fig. 2. (a) Diffusion coefficient D as a function of the mutation rate for $N = 10, 20, 30, 40, 80$ and $n = 10, 20, 30, 40, 80$ such that $N/n = 1$ after equilibration for 10^5 timesteps. Linear fitness matrix.
(b) Dependence of the ratio D/p on N/n.

the slope of

$$g(\tau) = \frac{1}{T_2 - T_1 + 1} \sum_{t=T_1}^{T_2} \|\mathbf{p}(t+\tau) - \mathbf{p}(t)\|^2 \tag{8}$$

over a suitable measurement interval $[T_1, T_2]$.

For the diffusion coefficients one finds empirically that the ratio D/p depends only on the ratio of N/n, see Figure 2a. Hence we plot the slope D/p as a function of N/n in Figure 2b. The data are consistent with an ansatz $D \propto p(N/n)^{-1.5\pm0.1}$. Our time unit is one simulation step. A physically more meaningful time unit would be the *generation time* in which on average every member of the population has been picked once for an attempt to replicate, i.e., $\tau' = \tau/N$. Hence the "physical" diffusion coefficient would be $D' = ND$.

4. Discussion

We have shown that in our model with a simple interaction matrix, the dynamics exhibits a quasispecies-like behavior. The simulations of Lindgren and Forst [5,9] suggest that the qualitative picture stays the same even with more complicated interaction matrices. Interactions where the fintess values a_{kj} are falling with distance, show the same qualitative result. (Data not shown). It would be interesting to look at the case of an RNA model. There, the phenotypes are equivalence classes of genotypes with parameters $a_{kj} = \Phi(f(\mathbb{I}_k), f(\mathbb{I}_j))$, where $f(\mathbb{I}_k)$ is the fitness of the phenotype belonging to sequence $\mathbb{I}_k$ and the map from genotypes to phenotypes is

many to one. A natural framework for such simulations is the RNA folding model [10].

Acknowledgements

This work was supported in part by the Austrian *Fonds zur Förderung der Wissenschaftlichen Forschung*, Project No. P-13887-MOB.

References

[1] Derrida B. and Peliti L. Evolution in a flat fitness landscape. *Bull. Math. Biol.*, **53**, 355–382 (1991).

[2] Eigen M. Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften*, **58**, 465–523 (1971).

[3] Eigen M., McCaskill J.S., and Schuster P. The molecular quasi-species. *Adv. Chem. Phys.*, **75**, 149–263 (1989).

[4] Forst C.V. Molecular evolution of catalysis. *J. Theor. Biol.*, **205**, 409–431 (2000).

[5] Forst C.V., Reidys C., and Weber J. Evolutionary dynamics and optimization: neutral networks as model-landscape for RNA secondary structure folding landscapes. In Advances in Artificial Life, *Lecture Notes in Artificial Intelligence **929***, F. Morán, A. Moreno, J.J. Merelo, and P. Chacón, eds., pp. 128–147 (Springer, Berlin, 1995).

[6] Gillespie D.T. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.*, **22**, 403 (1976).

[7] Higgs P.G. and Derrida B. Stochastic models for species formation in evolving populations. *J. Math. A: Math. Gen.*, **24**, L985–L991 (1991).

[8] Huynen M.A., Stadler P.F., and Fontana W. Smoothness within ruggedness: the role of neutrality in adaptation. *Proc. Natl. Acad. Sci. (USA)*, **93**, 397–401 (1996).

[9] Lindgren K. Evolutionary phenomena in simple dynamics. In *Artificial Life II*, pp. 295–312 (Addison Wesley, Reading, MA, 1992).

[10] Schuster P., Fontana W., Stadler P.F., and Hofacker I.L. From sequences to shapes and back: A case study in RNA secondary structures. *Proc. Roy. Soc. Lond. B*, **255**, 279–284 (1994).

[11] Schuster P. and Sigmund K. Replicator dynamics. *J. Theor. Biol.*, **100**, 533–538 (1983).

[12] Serva M. and Peliti L. A statistical model of an evolving population with sexual reproduction. *J. Math. A: Math. Gen.*, **24**, L705–L709 (1991).

[13] Stadler P.F. and Schuster P. Mutation in autocatalytic networks — an analysis based on perturbation theory. *J. Math. Biol.*, **30**, 597–631 (1992).