The error threshold in the quasispecies model as a phase transition in a 2-dimensional lattice model

Hanrong Chen
School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138

The quasispecies model is a simple model of molecular evolution, and features the existence of an "error threshold" beyond which the mutation rate is so high that all genetic information is lost. Following the work of Leuthausser, this model can be mapped to a problem in equilibrium statistical mechanics, and in particular, the error threshold is analogous to a phase transition occuring at the surface of a lattice model. Here, we review this mapping and the error thresholds for various simple replication landscapes.

I. INTRODUCTION

All forms of life use nucleic acids (either DNA or RNA) to carry genetic information that codes for proteins, including those that replicate the nucleic acids and enable the cell to reproduce. However, one might speculate that at the dawn of life, such complex machinery did not exist, and instead, nucleic acids were responsible for their own self-replication. Naively, in this picture "fitter" species (in the chemical, not biological, sense) are those that replicate more quickly. Motivated by the origin-of-life problem, a model of the evolution of such molecular replicators was introduced and studied by Manfred Eigen and others in the 1970s [1]. Indeed, for a variety of fitness, or replication, landscapes with low mutation rates, it was found that the stationary population would be localized at fitness peaks. Such populations consisting of similar molecular sequences are called "quasispecies". Interestingly, it was also found that the existence of these quasispecies is dependent on mutations; if the mutation rate is too high, beyond a certain "error threshold", no quasispecies survives, and all genetic information is lost.

It turns out that there is a correspondence between the quasispecies model and a 2-dimensional lattice model in thermodynamics equilibrium, and the error threshold becomes analogous to a phase transition in the surface layer [2]. This mapping allows us to use existing methods in statistical mechanics to analyze complicated replication landscapes. Here, we review this mapping, following the treatment in [2, 3]. We introduce 3 replication landscapes of increasing complexity, and present results finding the error thresholds in these cases.

II. FROM THE EVOLUTION EQUATION TO A LATTICE MODEL

In the quasispecies model, we consider sequences $(s_1, s_2, ..., s_L)$ of fixed length L where each variable s_l would represent one of the 4 different bases of DNA or RNA; here, we follow previous work [2, 3] and consider $s_l = \pm 1$, which might represent either a purine (A, G) or pyrimidine (C, T or U) base. Thus, there are 2^L possible sequences denoted by S_j, $j = 1...2^L$. In the infinite population, deterministic limit, the concentration of each sequence S_j, $x_j(t)$, obeys the following dynamics:

$$\dot{x_j}(t) = A[S_j]x_j - \sum_{k \neq j} W_{kj}x_j + \sum_{k \neq j} W_{jk}x_k \qquad (1)$$

where each $A[S_j]$ is the replication rate of sequence S_j, and the W_{kj} are the mutation rates from sequence S_j to S_k. The 2nd and 3rd terms on the right-hand side thus represent mutations away from, and to, sequence S_j, respectively. The dynamics of this equation is characterized by the eigenvalues and eigenvectors of the matrix W with entries W_{kj}, and in particular the steady-state, stationary population is represented by the eigenvector with the largest eigenvalue.

Now, we delineate how we can recast this model in statistical mechanics terms. Let us discretize time, such that the concentrations $x_k(t)$ of each S_k become $x_k(i)$ for generations $i = 0...n$. We may thus write the population of the nth generation as:

$$X(n) = W^n X(0) \qquad (2)$$

where $X(i) = (x_1(i), ..., x_{2^L}(i))$. W may actually be interpreted as the transfer matrix giving the probability of configuration $X(i+1)$ given the configuration of the nearest-neighbor sites $X(i)$. In this picture, we may describe the evolution of the population as a 2-dimensional lattice system, with variables $s_j(i) = \pm 1$ on rows $i = 0...n$, and $j = 1...L$ representing the position along the molecule. In general, the interaction between nearest-neighbor rows in the time direction is given by $h[S(i), S(i+1)]$ for an arbitrary function h, and thus the Hamiltonian for the system is:

$$H = \sum_{i=0}^{n-1} h[S(i), S(i+1)] \qquad (3)$$

The transfer matrix W is thus given by:

$$W_{jk} = \exp\left(-\beta h[S_j, S_k]\right) \qquad (4)$$

Thus, any replication matrix W gives us an effective Hamiltonian (3) of this lattice model representation. However, we consider the simplified case with random point mutations. Assuming that the fidelity of replication at each site in the sequence is independent of any other site, and has probability q, so that the probability of mutation per site is $1-q$, the mutation rate can be written as:

$$W_{jk} = A[S_k] q^{L-d_{jk}} (1-q)^{d_{jk}} \tag{5}$$

where d_{jk} is the number of inequivalent sites between sequences j and k:

$$d_{jk} = \frac{1}{2}\left(L - \sum_{k=1}^{L} s_j s_k\right) \tag{6}$$

Recall our definition of $A[S_k]$ as the replication landscape, and we will consider several specific examples of this. By comparing (4) with (5), we get the following effective Hamiltonian:

$$-\beta H = \sum_{i=0}^{n-1}\left(\beta \sum_{j=1}^{L} s_j(i)s_j(i+1) + \ln A_i\right) + \frac{nL}{2}\ln[q(1-q)] \tag{7}$$

where $\beta = \frac{1}{2}\ln[q/(1-q)]$, and for $\frac{1}{2} < q < 1$ (the case $0 < q < \frac{1}{2}$ is given by the mapping $\beta \to -\beta$). Note that we exclude the interactions in the final generation $i = n$, since in (4) the rate of replication only depends on the "mother" sequence. In conclusion, we have mapped the original dynamical equation to a lattice model with Hamiltonian (7), and finding the error threshold amounts to studying the critical properties of this model.

III. THE ERROR THRESHOLD FOR SIMPLE REPLICATION LANDSCAPES

To compute the steady state we take the $n \to \infty$ limit and get a semi-infinite lattice in the time direction, and a finite length L in the other. Our goal is to compute the error threshold for several replication landscapes $A[S]$ given below. It is far easier to analyze the critical properties of the bulk than the surface, but it has been found that the behavior near the threshold is strongly influenced by its surface character [3], so a full solution of the surface system may be required for understanding the nature of the transition. While the transfer matrix technique may be employed, it becomes computationally difficult for large L, and here we present the cluster-variational method as was used in [3].

A. Single sharp peak

The simplest and most well-studied example of a replication landscape is a flat one except for a single sharp peak corresponding to a "master sequence" $S_0 = (\xi_1, ..., \xi_L)$. The replication rates are then given by:

$$A[S_0] = A_0 \tag{8}$$
$$A[S] = A_1 < A_0, \; S \neq S_0 \tag{9}$$

The effective Hamiltonian (3) becomes:

$$-\beta H = \sum_{i=0}^{n-1}\left(\beta \sum_{j=1}^{L} s_j(i)s_j(i+1) + \ln\frac{A_0}{A_1}\prod_{j=1}^{L}\frac{\xi_j s_j(i)+1}{2}\right) \tag{10}$$

where we have dropped analytic terms that don't affect the transition. Figure 1 shows the result of the cluster-variational method [3]: the relative concentration of the sum of all the sequences at the same distance from the master sequence. Both show a clear error threshold beyond which the distribution is random over sequence space, but while the bulk transition appears discontinuous, the surface one is a smooth second-order one.

We may characterize this phase transition using an order parameter: a natural one to use is the projection of the population onto the master sequence S_0, given by:

$$m = \frac{1}{N}\sum_{j=1}^{L} \xi_j < s_j > \tag{11}$$

where the angular brackets denote the statistical average over the population in one generation. $m = 1$ is the case where the population exclusively consists of the master sequence, while $m = -1$ is that where the population exclusively consists of sequences complementary to the master sequence. $m = 0$ denotes a population of random sequences. In Figure 2, the order parameter is evaluated in the bulk m_b and at the surface m_s, and again we see a difference between the bulk and surface transitions.

B. Double-peak landscape

Next, let us consider a flat landscape with not one, but two peaks corresponding to $S_0 = (\xi_1, ..., \xi_L)$, as before, and its complement $S_L = (-\xi_1, ..., -\xi_L)$. If $A[S_0] = A[S_L]$, we have a degenerate quasispecies, and below the error threshold, the population is a mixture of both sequences. However, any small difference between $A[S_0]$ and $A[S_L]$ will break this degeneracy, and the steady-state population will only consist of one of them. However, a more interesting extension would be

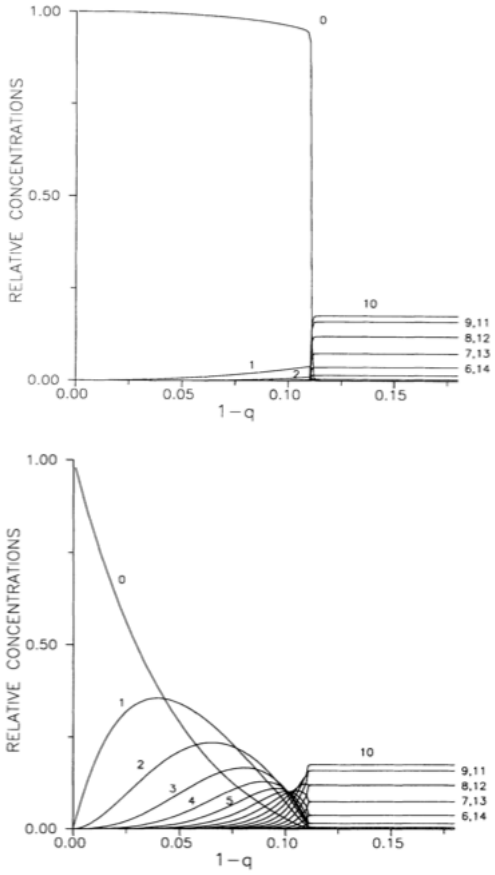

Figure 1: Concentration as a function of mutation rate $1-q$ for a single-peak landscape with $A_0/A_1 = 10$ and $L = 20$, as computed in the bulk (top) and surface (bottom).

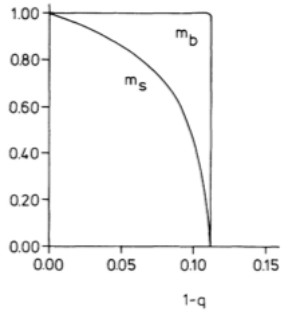

Figure 2: The order parameter for the bulk, m_b, and surface, m_s, for the single-peak landscape as in Figure 1.

to consider one peak with a high but narrow maximum and the other with a lower but broader one. An example of this is:

$$A[S_0] \;=\; A_0,\; A[S_L] = A_L,\; A[S_{L-1}] = A_{L-1} \quad (12)$$

and

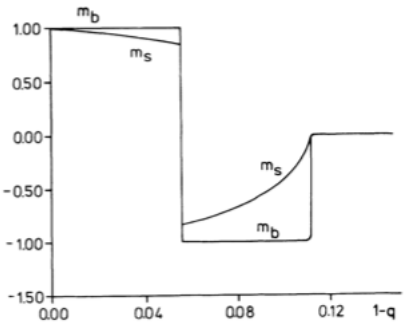

Figure 3: The order parameter for the bulk, m_b, and surface, m_s, for the double-peak landscape with $L = 20$, $A_0/A_1 = 10$, $A_{20}/A_1 = 9.9$ and $A_{19}/A_1 = 2$.

$$A[S] = A_1,\; S \neq S_0,\; S_L,\; S_{L-1} \quad (13)$$

where $A_0 > A_L > A_{L-1} > A_1$.

For low mutation rates, the quasispecies is localized at S_0, but as $1-q$ increases, the broader quasispecies will become favorable. Figure 3 shows the same order parameter (11), where $m = 1$ for S_0 and $m = -1$ for S_L. Both the bulk and surface distributions show a sharp transition from $m = 1$ to $m = -1$ near $1 - q = 0.06$, and there is as expected a smoother transition to $m = 0$, occurring at an error threshold $1 - q = 0.11$.

C. Mattis landscape

Our final example is the Mattis landscape, which like the sharp-peak landscape depends only on the distance to a master sequence $S_0 = (\xi_1, ..., \xi_L)$, but in a smooth way:

$$A[S] = \exp\left(\frac{K}{2L^2}\sum_{j\neq j'}^{L}\xi_j\xi_{j'}s_j s_{j'}\right) \quad (14)$$

The effective Hamiltonian (3) in this case is:

$$-\beta H = \sum_{i=0}^{n-1}\left(\beta\sum_{j=1}^{L}s_j(i)s_j(i+1) + \frac{K}{2L^2}\sum_{j\neq j'}^{L}\xi_j\xi_{j'}s_j(i)s_{j'}(i)\right) \quad (15)$$

Note that as $L \to \infty$, (14) gives $A_{max}/A_{min} \approx \exp(K/2)$. The results are shown in Figure 4.

IV. DISCUSSION

We have reviewed an interesting correspondence between a model of molecular evolution and a lattice model,

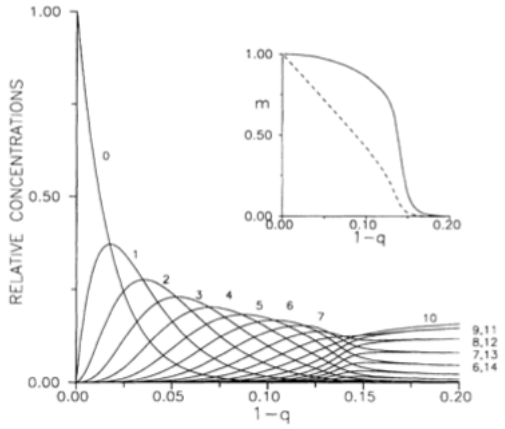

Figure 4: Relative concentrations and the order parameter versus the mutation rate for the Mattis landscape.

in particular showing how the "error threshold" may be computed from the critical temperature of a phase transition. As demonstrated in [3], one must be careful in analyzing this transition, which typically shows different behavior in the bulk and on the surface. In particular, the surface distribution changes much more smoothly than that of the bulk near criticality.

Finally, we have so far only been concerned with the deterministic, infinite population limit, which corresponds to perfect thermodynamic equilibrium in our lattice model. It is known that the error threshold exists for simple landscapes, even for finite populations [1]. Interesting to check would be how these results carry over to finite populations where stochastic effects are strong for the more complicated landscapes.

[1] M. Eigen, Naturwissenschaften 58, 465 (1971). M. Eigen and P. Schuster, Naturwissenschaften 64, 541 (1977); 65, 7 (1978); 65, 341 (1978).
[2] I. Leuthausser, J. Stat. Phys. 48, 343 (1987).
[3] P. Tarazona, Phys. Rev. A 45, 6038 (1992).