Error Threshold in Eigen's Quasispecies Model of Molecular Evolution

Alexander Bratus

Moscow State University, Russia

Yuri Semenov

Moscow State University of Railway Engineering, Russia

Artem Novozhilov

North Dakota State University, USA

Moscow, 2016

Quasispecies theory:

Manfred Eigen, born 1927

- M. Eigen, Naturwisenschaften, 58(10), 1971:465–523
- M. Eigen, P. Schuster, The Hypercycle, Springer, 1979
- M. Eigen, J. McCaskill, P. Schuster, J Phys Chem, 92(24), 1988:6881-6891

DIE NATURWISSENSCHAFTEN

58. Jahrgang, 1971 Heft 10 Oktober

Selforganization of Matter and the Evolution of Biological Macromolecules

MANFRED EIGEN*

Max-Planck-Institut für Biophysikalische Chemie,
Karl-Friedrich-Bonhoeffer-Institut, Göttingen-Nikolausberg

Mathematical side of the story:

The variables of the dynamical system are the concentrations of individual polynucleotide sequences: $[I_i] = c_i(t)$. We are interested, essentially, in the relative concentrations of the different species

$$x_i(t) = c_i(t) / \sum_{i=1}^{n} c_i(t); \quad i = 1, 2, \ldots, n \tag{12}$$

The resulting kinetic equations, around which quasi-species theory centers, are then

$$\mathrm{d}x_i(t)/\mathrm{d}t \equiv \dot{x}_i(t) = (W_{ii} - \bar{E}(t))x_i(t) + \sum_{k \neq i} W_{ik} x_k(t); \quad i, k = 1, 2, \ldots, n \tag{13}$$

The mean excess production

$$\bar{E}(t) = \sum_{i=1}^{n} x_i(t) E_i \tag{14}$$

of the population may be physically compensated by a dilution

Ref: M. Eigen, J. McCaskill, P. Schuster, J Phys Chem, 92(24), 1988: 6881-6891

Model statement:

Consider a very large (infinite) population of haploid individuals (sequences) with fixed genome length N composed of two-letter alphabet, say, $\{0, 1\}$, therefore 2^N different sequences. For example, if $N = 4$ then it is possible to have 2^4 different sequences:

$$
\begin{array}{cccc}
[0000], & [0001], & [0010], & [0011], \\
[0100], & [0101], & [0110], & [0111], \\
[1000], & [1001], & [1010], & [1011], \\
[1100], & [1101], & [1110], & [1111].
\end{array}
$$

Different sequences have different fitnesses (selection) and it is possible that 0 at any site mutates into 1 and vice versa (mutations).

Eigen's quasispecies model:

The quasispecies model takes into account two evolutionary forces: selection and mutations. The selection can be described by the diagonal matrix of fitnesses

$$\boldsymbol{W} = \mathrm{diag}(w_1, \ldots, w_l),$$

and mutations are described by the double stochastic matrix

$$\boldsymbol{Q} = [q_{ij}]_{l \times l},$$

where q_{ij} is the probability that macromolecule j produces macromolecule i, which can be further defined as

$$q_{ij} = q^{N - d_{ij}} (1 - q)^{d_{ij}},$$

where q is the fidelity of replication per site (i.e., $1 - q$ is the probability of mutation per site), and d_{ij} is the Hamming distance between sequences i and j.

Eigen's quasispecies model:

Let $n_j(t)$ be the number of sequences of type j at time t. Then, using the notations from the previous slide,

$$\dot{n}_j = \sum_{i=1}^{l} q_{ji} w_i n_i, \quad j = 1, \ldots, 2^N.$$

Switching from the absolute sizes to the frequencies, $p_j(t) = \frac{n_j(t)}{\sum_i n_i(t)}$, leads to

$$\dot{p}_j = \sum_{i=1}^{l} q_{ji} w_i p_i - \bar{w}(t) p_j, \quad j = 1, \ldots, 2^N,$$

or, in the matrix form,

$$\dot{\boldsymbol{p}} = \boldsymbol{Q}\boldsymbol{W}\boldsymbol{p} - \bar{w}(t)\boldsymbol{p}.$$

Here $\bar{w}(t)$ is the mean population fitness,

$$\bar{w}(t) = \sum_{j=1}^{l} w_j p_j(t) = \boldsymbol{w} \cdot \boldsymbol{p}(t).$$

Eigen's quasispecies model:

The only equilibrium point of the Eigen's quasispecies model satisfies the equation

$$\boldsymbol{QW}\boldsymbol{p} = \bar{w}\boldsymbol{p},$$

which is the eigenvalue problem for matrix $\boldsymbol{QW}$. It it quite straightforward to show that, according to the Perron–Frobenius theorem, there is always a solution to this problem, where $\bar{w}$ is the leading (dominant) real positive eigenvalue and $\boldsymbol{p} > 0$ is a corresponding eigenvector. Moreover, this equilibrium point is globally asymptotically stable.
This vector $\boldsymbol{p}$ was called the *quasispecies* by Eigen.

Crow–Kimura quasispecies model:

In the following we consider a modification of the original Eigen's model. This modification takes into account two things:

- First, we assume that all the sequences with the same number of 1s have exactly the same fitness. This means that we do not distinguish, e.g., between sequences [0010] and [0100] and thus have to deal with $N+1$ classes of sequences instead of 2^N types of macromolecules.
- Second, instead of taking into account *probabilities*, as was done originally by Eigen, we will concentrate, as it is more natural in the continuous time settings, on the *rates* μ_{ij}.

The evolutionary force of *selection* is included through the fitness landscape, which in our case is given by a diagonal matrix

$$\boldsymbol{M} = \mathrm{diag}(m_0, \ldots, m_N),$$

where m_j is the fitness of the j-th class of sequences.

Model statement:

The second evolutionary force is *mutation*.

In particular, assuming $N+1$ classes of sequences, we have that the mutations μ_{ij} (i.e., the mutation rate from class j to class i) can be described by the matrix

$$\boldsymbol{\mathcal{M}} = (\mu_{ij}) = \mu\boldsymbol{Q} = \mu \begin{bmatrix} -N & 1 & 0 & 0 & \dots & \dots & 0 \\ N & -N & 2 & 0 & \dots & \dots & 0 \\ 0 & N-1 & -N & 3 & \dots & \dots & 0 \\ 0 & 0 & N-2 & -N & \dots & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \dots & 2 & -N & N \\ 0 & 0 & \dots & \dots & 0 & 1 & -N \end{bmatrix},$$

where μ is the mutation rate per site per sequence per replication event.

Model statement:

Let $\boldsymbol{p}(t)$ denote the vector of frequencies of different classes of sequences, then, assuming uncoupled reproduction and mutation events, we arrive at

$$\dot{\boldsymbol{p}}(t) = (\boldsymbol{M} + \mu\boldsymbol{Q})\boldsymbol{p}(t) - \overline{m}(t)\boldsymbol{p}(t),$$

where

$$\overline{m}(t) = \boldsymbol{m} \cdot \boldsymbol{p}(t) = \sum_{i=0}^{N} m_i p_i(t)$$

is the mean population fitness.

This model is often called a paramuse of Crow–Kimura quasispecies model with permutation invariant fitness landscape.

Ref: Baake and Gabriel, Annual Reviews of Computational Physics VII, 1999: 203–264

Ref: Crow and Kimura, An introduction to population genetics theory, 1970

Elementary results:

The asymptotic behavior of the quasispecies model is determined by the equilibrium $\boldsymbol{p} = \lim_{t\to\infty} \boldsymbol{p}(t)$, which solves the eigenvalue problem

$$(\boldsymbol{M} + \mu\boldsymbol{Q})\boldsymbol{p} = \overline{m}\,\boldsymbol{p},$$

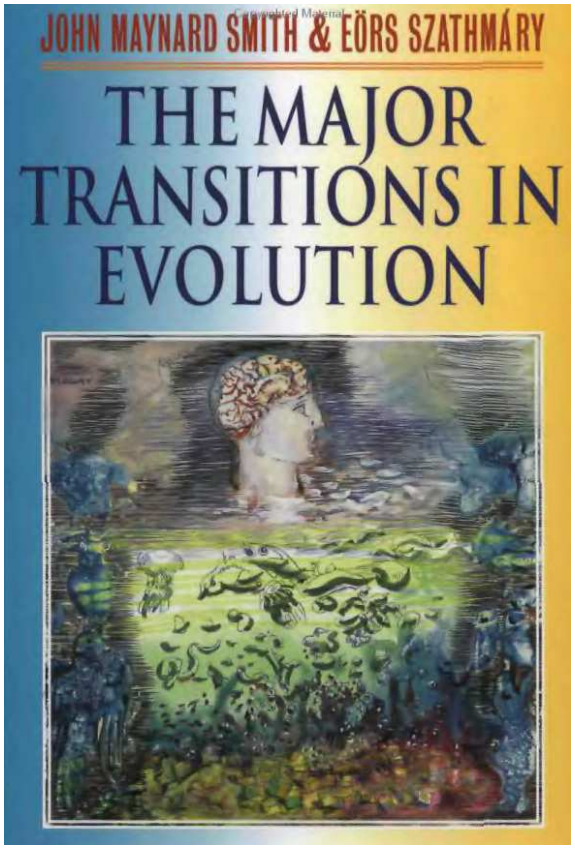
where

$$\overline{m} = \boldsymbol{m}\cdot\boldsymbol{p} = \sum_{i=0}^{N} m_i p_i.$$

By Perron–Frobenius theorem it follows that there is a unique positive solution $\boldsymbol{p} > 0$, which is the right eigenvector of $\boldsymbol{M} + \mu\boldsymbol{Q}$ corresponding to the simple real dominant eigenvalue $\lambda = \overline{m}$.

This vector $\boldsymbol{p}$ was called by Eigen the *quasispecies*. It is globally stable for the quasispecies system. We are mostly interested in properties of $\overline{m}$ and $\boldsymbol{p}$ depending on the fitness landscape $\boldsymbol{M}$ and mutation rate μ, therefore, we use the notation $\overline{m} = \overline{m}(\mu)$ and $\boldsymbol{p} = \boldsymbol{p}(\mu)$ for the mean fitness and equilibrium distribution.

Known results: The error threshold



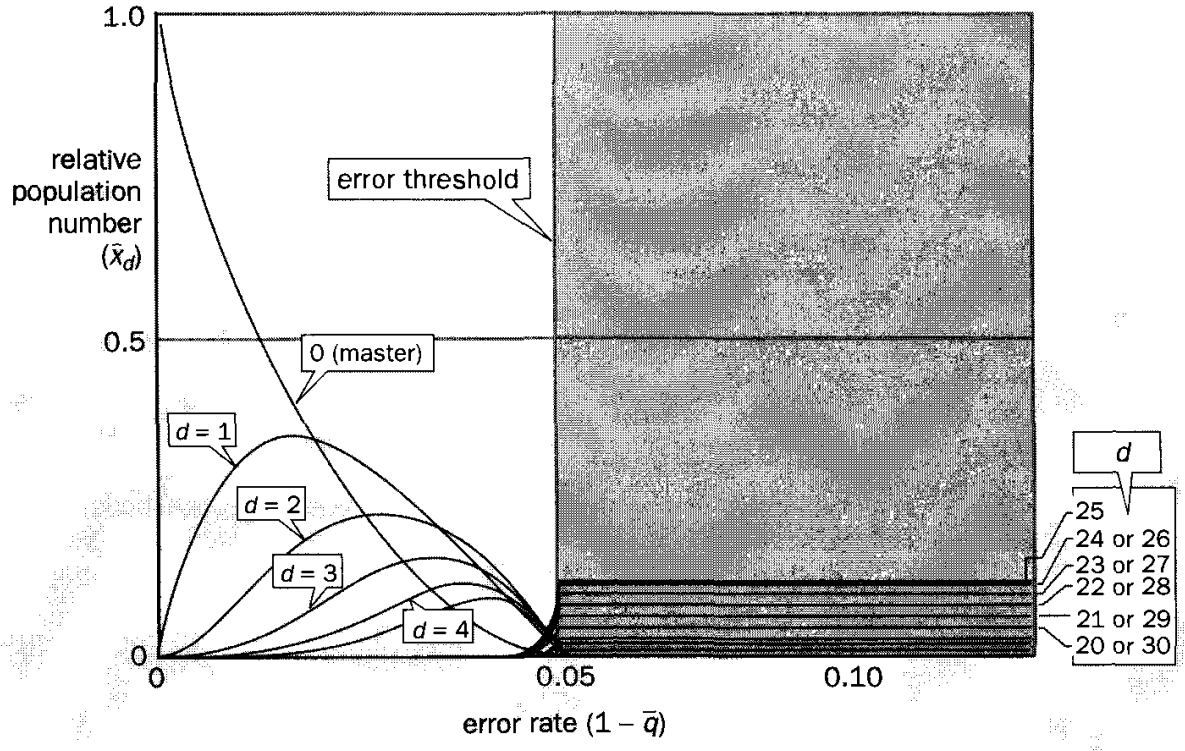
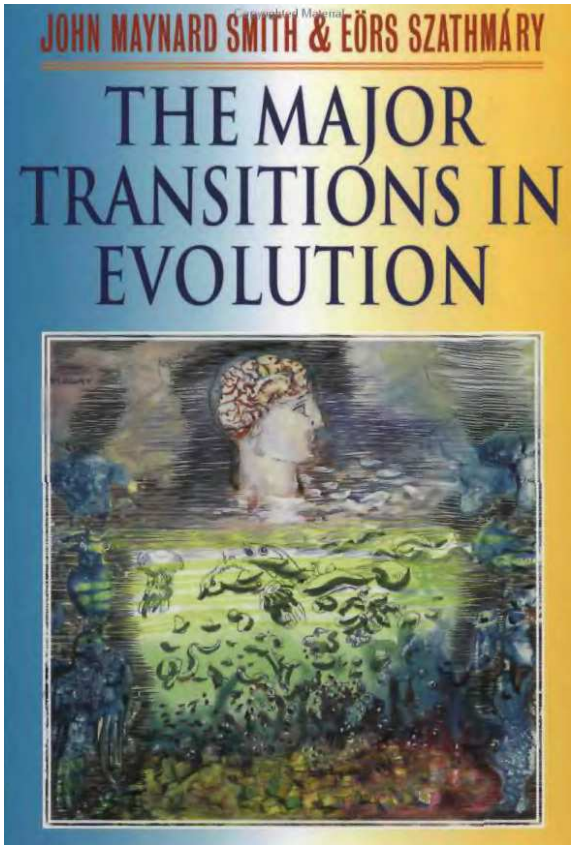


Fig. 4.5 The error threshold (Swetina & Schuster, 1982). d is the number of differences between a particular sequence and the 'master' sequence that has the highest fitness. Thus the curve for $d = 3$ is the proportion of sequences differing from the master by exactly three mutations. In this simulation, the sequence was of 50 sites, each occupied by one of two kinds of base. The fitnesses of all sequences other than the master were equal. Beyond the threshold, all sequences were equally frequent. Since the numbers of sequences with $d = 23$ and $d = 27$, for example, are the same, the frequencies of the two classes were also equal, as shown here.

Known results: The error threshold



We are interested in the coexistence of master and mutants. The alternative is that only mutants are present at equilibrium: selection is not able to maintain the master sequence against mutational decay. We therefore set both rates of change to zero, and ask whether an equilibrium with non-zero x_m can exist. It turns out that such an equilibrium requires

$$Q > A_j/A_m = 1/s, \tag{4.4}$$

where s is the selective superiority of the master. We know, however, that

$$Q = q^N \approx e^{-N(1-q)}. \tag{4.5}$$

Combining the last two equations we have

$$N < \ln s/(1-q), \tag{4.6}$$

which means that the selectively maintainable amount of information (N) is limited by the copying fidelity per digit (q) (Eigen, 1971).

Known results: The error threshold

Opinion

Open Access

The fundamental units, processes and patterns of evolution, and the Tree of Life conundrum

Eugene V Koonin* and Yuri I Wolf

It is almost as intuitively clear that, although for evolution to occur, replication must be error-prone (and, replication is, in any case, error-prone owing to physical constraints), there must also exist an error threshold such that an above-threshold error rate renders evolution impossible. Extrapolating to the extreme (absurd), it is obvious that a "replication" process that incorporates nucleotides randomly is not conducive to evolution (and, of course, does not really qualify as replication). Spiegelman's experiments stimulated theoretical work by Eigen and coworkers that put the link between replication and evolution into a mathematical framework and quantified the requirements to the replication error rate [30]. Eigen's seminal work and subsequent, increasingly sophisticated analysis showed that the error threshold, that is, the minimal fidelity that is required for mutations to be fixed and, accordingly, for evolution to proceed, is relatively low, in the range of 1-10 errors per replication cycle (the exact number remains a matter of debate) [31-34]. It appears that most if not all replicating entities exist on or close to the edge of the "Eigen cliff", with the fidelity of replication only slightly exceeding the minimal requirement (Figure 1) [35].

An intriguing question is whether evolution involves "selection for evolvability" [36,37] or the existence near the edge results from the opportunistic character of the evolutionary process whereby fidelity is increased to the extent strictly necessary but not far beyond that because further increase would incur substantial cost of selection. However, discussion of this important problem is beyond the scope of this article.

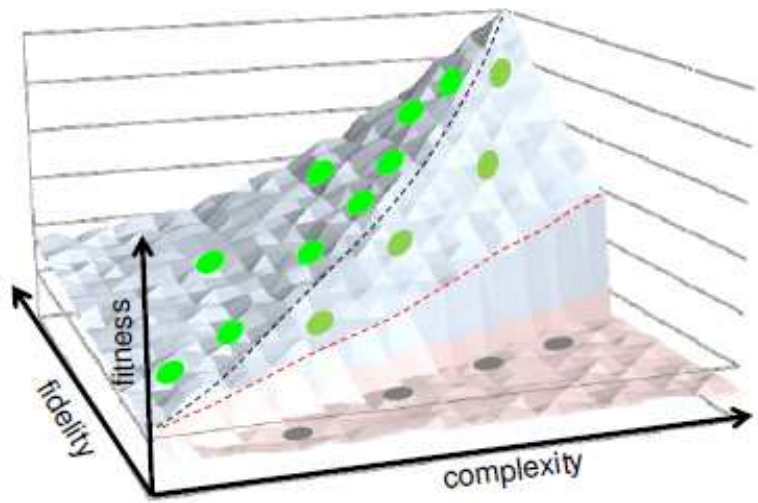


Figure 1
Replicating genetic elements exist close to the replication error threshold.

Known results: The error threshold

Ref: Swetina and Schuster, Bioph Chem, 16: 329–345, 1982

Consider the single peaked fitness landscape

$$\boldsymbol{M} = \mathrm{diag}(m_0, 0, \ldots, 0), \quad m_0 > 0.$$

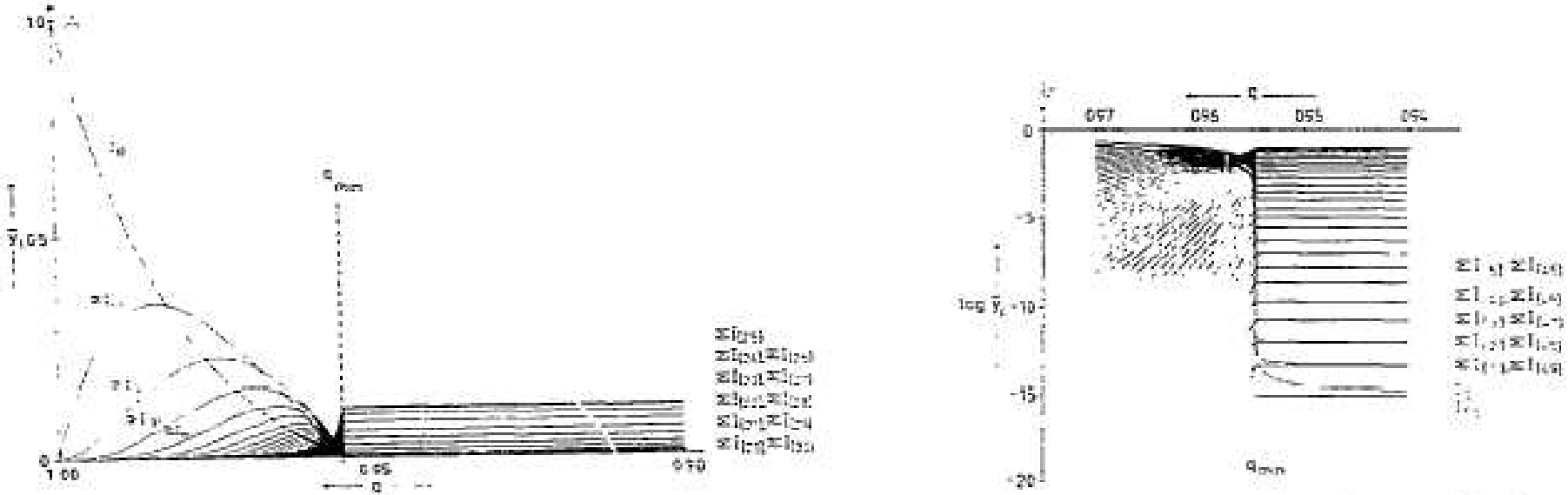


Fig. 10. Distribution of mutant classes as a function of the single-digit accuracy q for $\nu = 30$. Note the sharpness of the transition from direct to stochastic replication around q_{min}. This is seen best on the logarithmic plot. In the domain of stochastic replication individual concentrations become exceedingly small: $\bar{x}_i = 8.9 \times 10^{-16}$, $i = 0, \ldots 2^{30} - 1$. For basic definitions and numerical values see fig. 7.

Known results: The error threshold

Consider the single peaked fitness landscape

$$\boldsymbol{M} = \mathrm{diag}(m_0, m_1, \ldots, m_1), \quad m_0 > m_1.$$

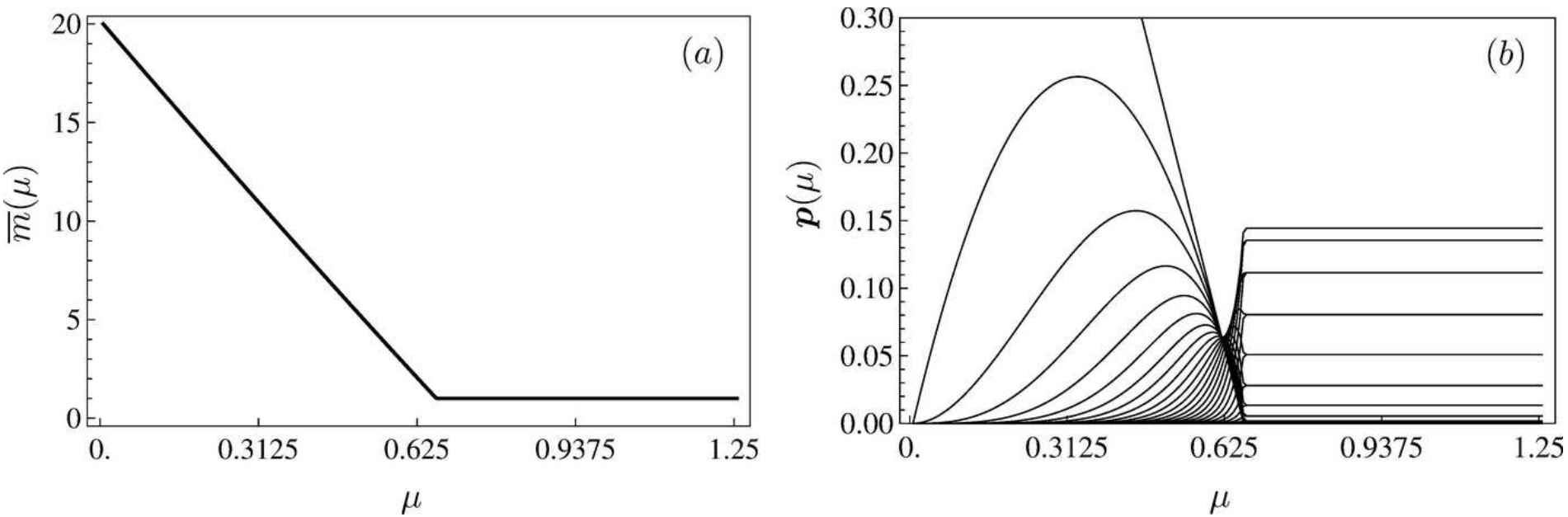


Figure : Error threshold in the quasispecies model with the single peaked fitness landscape ($\boldsymbol{m} = (m_0, m_1, \ldots, m_1)$, $m_0 > m_1$). The parameters are $N = 30$, $m_0 = 20$, $m_1 = 1$. (a) The mean population fitness $\overline{m}(\mu)$ versus the mutation rate; (b) the stationary quasispecies distribution versus the mutation rate

Known results: Statistical Physics

- Leuthäusser, I. (1986). An exact correspondence between Eigen's evolution model and a two-dimensional Ising system. The Journal of Chemical Physics, 84(3), 1884-1885.
- Tarazona, P. (1992). Error thresholds for molecular quasispecies as phase transitions: From simple landscapes to spin-glass models. Physical Review A, 45(8), 6038.
- Baake, E., Baake, M., & Wagner, H. (1997). Ising quantum chain is equivalent to a model of biological evolution. Physical Review Letters, 78(3), 559-562.
- Galluccio, S. (1997). Exact solution of the quasispecies model in a sharply peaked fitness landscape. Physical Review E, 56(4), 4526.
- Baake, E., & Wagner, H. (2001). Mutation-selection models solved exactly with methods of statistical mechanics. Genetical Research, 78(01), 93-117.
- Saakian, D. B., & Hu, C. K. (2006). Exact solution of the Eigen model with general fitness functions and degradation rates. Proceedings of the National Academy of Sciences of the USA, 103(13), 4935-4939.

Main idea:

We consider the eigenvalue problem

$$(\boldsymbol{M} + \mu \boldsymbol{Q})\boldsymbol{p} = \overline{m}\,\boldsymbol{p}, \quad \overline{m} = \boldsymbol{m} \cdot \boldsymbol{p},$$

where $\boldsymbol{p} = \boldsymbol{p}(\mu)$, $\overline{m} = \overline{m}(\mu)$ with a fixed fitness landscape $\boldsymbol{m}$.

We claim that this problem simplifies in the coordinates of the basis composed of the eigenvectors of the matrix $\boldsymbol{Q} = \boldsymbol{Q}_N$. Recall that

$$\boldsymbol{M} = \mathrm{diag}(m_0, \ldots, m_N), \qquad \boldsymbol{Q}_N = \begin{bmatrix} -N & 1 & 0 & \ldots & \ldots & 0 \\ N & -N & 2 & \ldots & \ldots & 0 \\ 0 & N-1 & -N & \ldots & \ldots & 0 \\ 0 & 0 & N-2 & \ldots & \ldots & 0 \\ \ldots & \ldots\ldots & \ldots & \ldots & \ldots & \\ 0 & 0 & \ldots & 0 & 1 & -N \end{bmatrix}$$

Proposition: For the matrix $\boldsymbol{Q} = \boldsymbol{Q}_N$:

1. The eigenvalues of $\boldsymbol{Q}_N$ are simple (all have algebraic multiplicities one) and given by
$$q_k = -2k, \quad k = 0, \ldots, N.$$

2. Let $\boldsymbol{v}_k^\top = (c_{0k}, \ldots, c_{Nk})$ be the right eigenvector of $\boldsymbol{Q}_N$ corresponding to q_k and normalized such that $c_{0k} = 1$, $\boldsymbol{C} = \boldsymbol{C}_N = (c_{ik})_{(N+1)\times(N+1)}$ be the matrix composed of $\boldsymbol{v}_k$ ($\boldsymbol{v}_k$ is the k-th column of $\boldsymbol{C}_N$). Then the generating function for the elements of the k-th column has the form
$$P_k(t) = \sum_{i=0}^{N} c_{ik}\, t^i = (1-t)^k (1+t)^{N-k}, \quad k = 0, \ldots, N.$$

3. $\boldsymbol{C}^2 = 2^N \boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix, or, equivalently,
$$\boldsymbol{C}^{-1} = 2^{-N} \boldsymbol{C}.$$

4. 1-norm of $\boldsymbol{C}$ is
$$\|\boldsymbol{C}\|_1 = \max_{0 \le k \le N} \sum_{i=0}^{N} |c_{ik}| = 2^N.$$

Main theorem:

There exists the limit

$$\bar{m}^* = \lim_{\mu \to \infty} \bar{m}(\mu) = 1 + \sum_{k=0}^{N} \frac{m_k - 1}{2^N} C_N^k$$

and the corresponding eigenvector

$$\boldsymbol{p}^* = \lim_{\mu \to \infty} \boldsymbol{p}(\mu) = 2^{-N} (C_N^0, C_N^1, \ldots, C_N^N).$$

Moreover,

$$\|\boldsymbol{p}(\mu) - \boldsymbol{p}^*\| \leq \frac{1}{\mu} \|\boldsymbol{M}\|_1.$$

Perturbation approximation:

$$(\boldsymbol{M} + \mu\boldsymbol{Q})\boldsymbol{p} = \bar{m}(\mu)\boldsymbol{p},$$

$$\bar{m}(\mu) = \bar{m}_0 + \bar{m}'(\mu_0)\Delta\mu + \frac{1}{2}\bar{m}''(\mu_0)(\Delta\mu)^2 + o((\Delta\mu)^2),$$

$$\boldsymbol{p}(\mu) = \boldsymbol{p}^0 + \boldsymbol{p}^1\Delta\mu + \boldsymbol{p}^2(\Delta\mu)^2 + o((\Delta\mu)^2),$$

$$\bar{m}'(0) = -N, \quad \bar{m}''(0) = \frac{N}{m_0 - m_1}.$$

Fix $\varepsilon > 0$. Call μ_ε ε-critical value if $|\bar{m}_{lim} - \bar{m}(\mu)| < \varepsilon$ for $\mu > \mu_\varepsilon$.

$$\mu_\varepsilon = (m_0 - m_1)\left(1 - \sqrt{1 - \frac{2(m_0 - 1)}{m_0 - m_1}N}\right)$$

if $\delta = \sum_{k=0}^{N} \frac{m_k - 1}{2^N} C_N^k < \varepsilon$, and

$$\mu_\varepsilon = (m_0 - m_1)\left(1 - \sqrt{1 - \frac{2(m_0 - 1)(\delta + 1)}{m_0 - m_1}N}\right), \quad \delta > \varepsilon.$$

The error threshold

Consider again the single peaked fitness landscape
$\boldsymbol{M} = \mathrm{diag}(m_0, m_1, \ldots, m_1), \quad m_0 > m_1.$

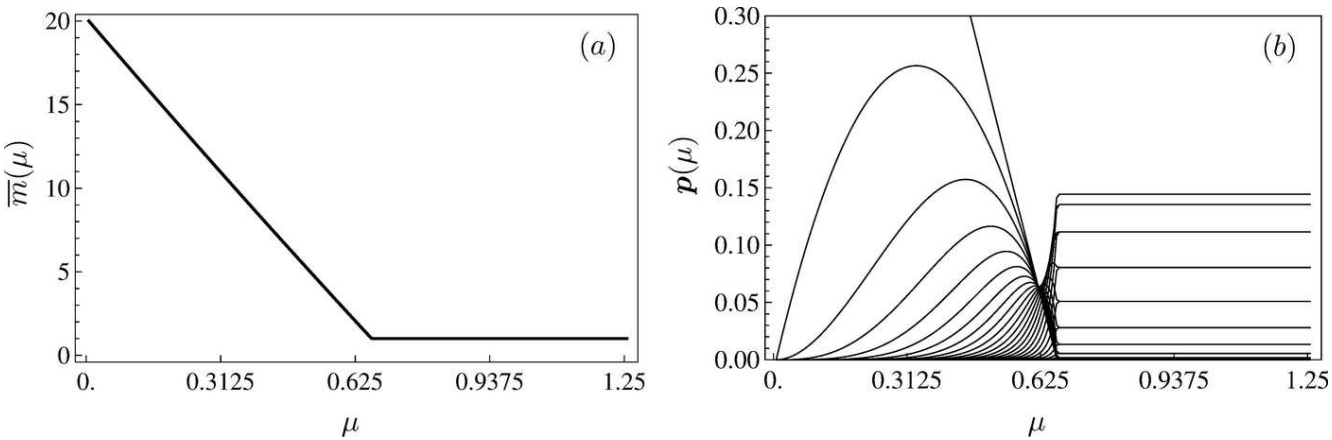


Figure : Error threshold in the quasispecies model with the single peaked fitness landscape ($\boldsymbol{m} = (m_0, m_1, \ldots, m_1)$, $m_0 > m_1$). The parameters are $N = 30$, $m_0 = 20$, $m_1 = 1$. (a) The mean population fitness $\overline{m}(\mu)$ versus the mutation rate; (b) the stationary quasispecies distribution versus the mutation rate

Here

$$\mu_{Eigen} = 0.633, \quad \mu_{\varepsilon} = 0.656,$$

whereas numerical computations suggest $\mu_{Error} = 0.66$.

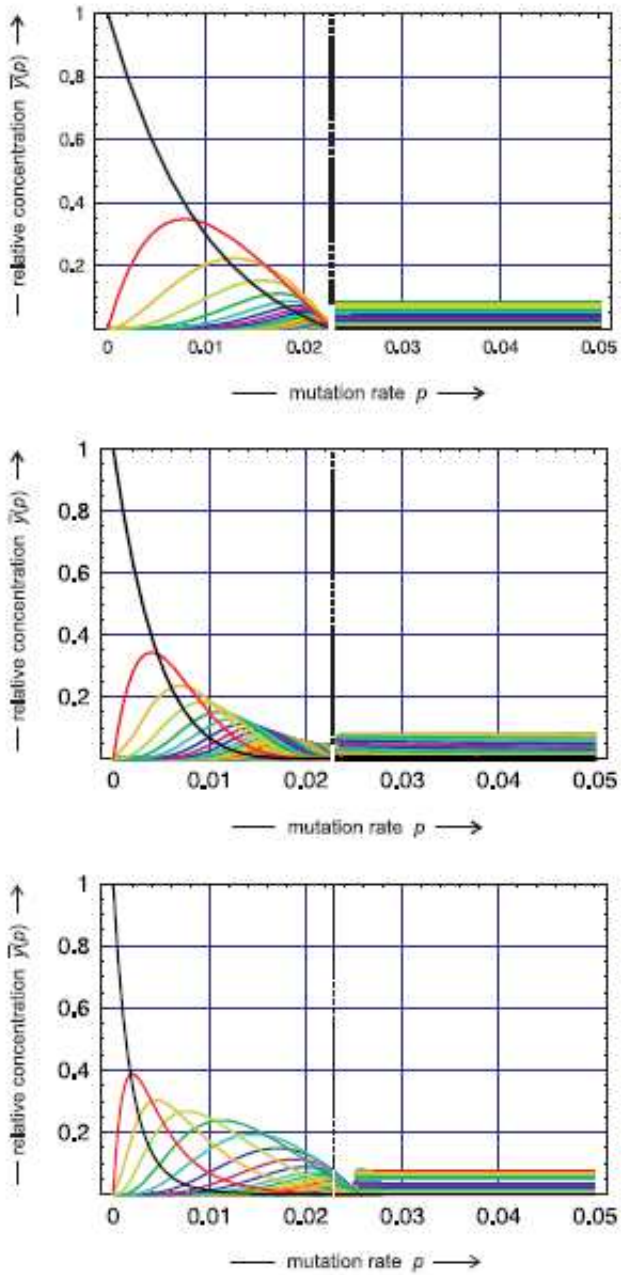


Figure 12: Error thresholds on different model landscapes. The figures show stationary concentrations of mutant classes as functions of the error rate, $\bar{y}_k(p)$, for sequences of chain length $\nu = 100$ with $f_0 = 10$ and $\overline{f}_{-0} = 1$ on three different model landscapes: the single peak landscape (upper part, $f =$ 1), the

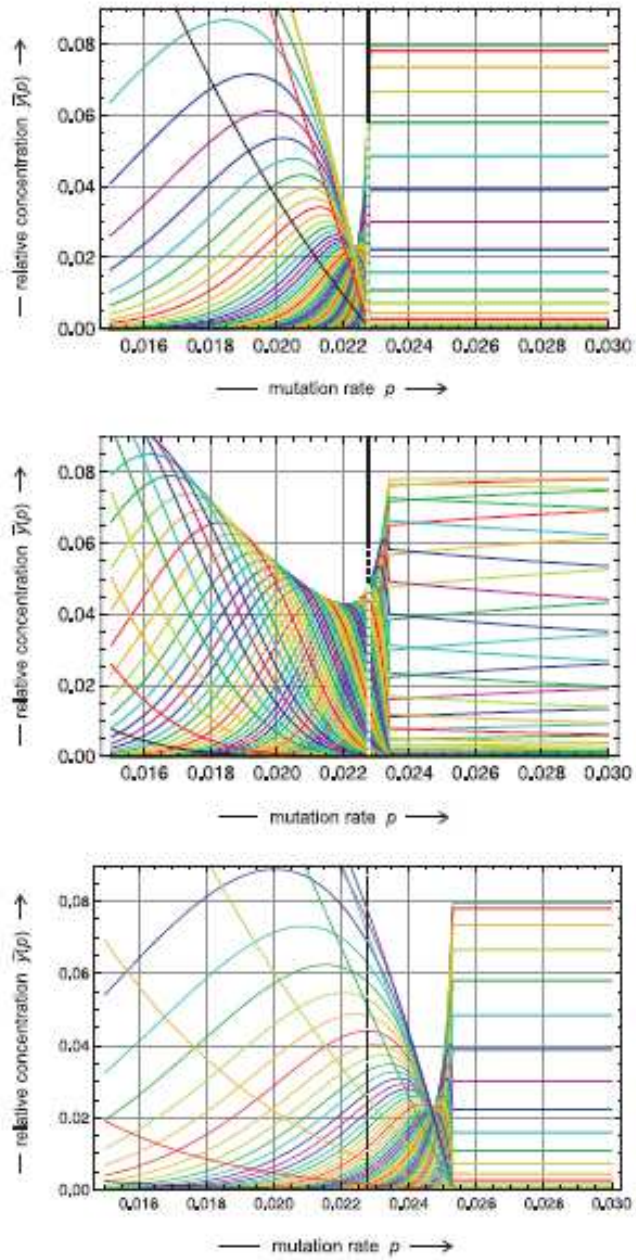


Figure 13: Error thresholds on different model landscapes. The three figures are enlargements of the plots from in figure 12. Stationary concentrations of mutant classes, $\bar{y}_k(p)$, are shown for the single peak landscape (upper part), the hyperbolic landscape (middle part), and the step-linear landscape (lower part; see

Thank you for your attention!

References:

- Bratus, A. S., Novozhilov, A. S., & Semenov, Y. S. (2014). Linear algebra of the permutation invariant Crow–Kimura model of prebiotic evolution. Math Biosciences
- Semenov, Y. S., Bratus, A. S., & Novozhilov, A. S. (2014). On the behavior of the leading eigenvalue of the Eigen evolutionary matrices. Math Biosciences,
- Semenov, Y. S., & Novozhilov, A. S. (2015) Exact solutions for the selection-mutation equilibrium in the Crow-Kimura evolutionary model. Math Biosciences
- Semenov, Y. S., & Novozhilov, A. S. (2016) On Eigen's quasispecies model, two-valued fitness landscapes, and isometry groups acting on finite metric spaces, Bulletin of Mathematical Biology