Mānoa Mini-Symposium on Physics of Adaptive Computation

How Much Information Can Natural Selection Maintain?

January 7, 2019

Lee Altenberg
altenber@hawaii.edu
http://dynamics.org/Altenberg/
Fundamental Processes in Biology: (1) Growth

**GROWTH**

**DECAY (NEGATIVE GROWTH)**

**MATHEMATICAL FORMS e.g.:**

\[ x(t+1) = \lambda x(t) \]

\[ \frac{d}{dt} x(t) = \lambda x(t) \]
**Fundamental Processes in Biology: (2) Transformation**

**TRANSFORMATION (i.e. CHANGE OF STATE)**

**MATHEMATICAL FORMS e.g.:**

\[ x(t+1) = \sum_y P_{xy} \ y(t) \]

\[ \frac{\partial}{\partial t} u(x, t) = \Delta u(x, t) \]
Growth and Transformation Combined

**GROWTH + TRANSFORMATION**

 MATHEMATICAL FORMS e.g.:

\[ x(t+1) = \sum_y P_{xy} \lambda_y y(t) \]

\[ \frac{\partial}{\partial t} u(x, t) = \Delta_x u(x, t) + \lambda(u) u(x, t) \]
Growth is a concentrating operator
Transformation is a *dispersing* operator
Growth and Transformation combined create a *search* operator
Basic question regarding the information stored by organisms:


What is the relationship between

1. mutation,
2. natural selection, and
3. the accumulation of information in the genome?
“A quasi-species is defined as a given distribution of macromolecular species with closely interrelated sequences, dominated by one or several (degenerate) master copies. . . .

Most important for Darwinian behavior are the criteria for internal stability of the quasi-species.
“If these criteria are violated, the information stored in the nucleotide sequence of the master copy will disintegrate irreversibly leading to an error catastrophe.

As a consequence, selection and evolution of RNA or DNA molecules is limited with respect to the amount of information that can be stored in a single replicative unit.”
Eigen and Schuster (1977, p. 555):

- “There is a threshold-relationship for the rate of mutation, at which evolution is fastest, but which must not be surpassed unless all the information thus far accumulated in the evolutionary process is to be lost.”

- “The number of molecular symbols of a self-reproducible unit is restricted, the limit being inversely proportional to the average error rate per symbol,” $p$. 
Li et al. (2015) Statistical properties and error threshold of quasispecies on single-peak Gaussian-distributed fitness landscapes

The results show that for different values of the fluctuation strength, the random fitnesses follow the Gaussian distribution. The relative concentration of the master sequence melts gradually in the deterministic Eigen model. In the random Eigen model, the averaged relative concentration becomes wider as the fluctuation strength increases. The upper limit of the crossover region obviously surpasses the error threshold given by the deterministic Eigen model. This fact should be considered when dealing with the practical problems of species evolution. Although the error threshold changes significantly, it becomes larger and larger. That is to say, the width of the error threshold increases with the fluctuation strength.

The relative concentration values for a given mutation rate can be understood as follows. The randomization becomes wider as the fluctuation strength increases. The relative difference in the relative concentration of two complementary classes get together because of the same degeneracy. The relative concentration of the master class, and 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 in the figure represents the master class, and 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 in the random Eigen model with 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0. (b) and (c) those in the random Eigen model with 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0. The crossover region is the error threshold given by the deterministic Eigen model. And the end point is the position of the crossover region.

The error threshold for single-peak fitness landscapes the evolution process of the population. Here complementary classes are the two classes which is a sharp point similar to a phase transition in physics. Over the error threshold, the sequences have negligibly small concentrations. The master sequence melts gradually in the deterministic Eigen model with a different mutant class. In the deterministic Eigen model with a different mutant class.

The relative concentration values for a given mutation rate can be understood as follows. The randomization becomes wider as the fluctuation strength increases. The relative difference in the relative concentration of two complementary classes get together because of the same degeneracy. The relative concentration of the master class, and 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 in the figure represents the master class, and 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 in the random Eigen model with 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0. (b) and (c) those in the random Eigen model with 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0. The crossover region is the error threshold given by the deterministic Eigen model. And the end point is the position of the crossover region.

The relative concentration values for a given mutation rate can be understood as follows. The randomization becomes wider as the fluctuation strength increases. The relative difference in the relative concentration of two complementary classes get together because of the same degeneracy. The relative concentration of the master class, and 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 in the figure represents the master class, and 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 in the random Eigen model with 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0. (b) and (c) those in the random Eigen model with 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0. The crossover region is the error threshold given by the deterministic Eigen model. And the end point is the position of the crossover region.
Papers per year with “error catastrophe” or “error threshold”
AND (quasispecies OR Eigen)

20 years: 1 to 15 citations per year
Four distinct threshold phenomena from increasing mutation rates:

1. “A kink in the population mean fitness,
2. the loss of the wildtype from the population,
3. complete mutational degradation [error catastrophe],
4. a jump in the population mean of the mutational distance”

In Eigen and Schuster’s fitness landscape, these four “error thresholds” happen to coincide. But they may not even exist in the general case.
The Lore: These insights and caveats still have not penetrated the literature


“Eigen and Schuster (1977) showed that the limit in question is indeed a universal necessity because it is a consequence of fundamental theorems that apply to all self-replicating systems.

The maximum length of the molecules is determined by the replication errors that are inevitably present in any replication process, because beyond that limit the system is overtaken by a runaway error catastrophe and collapses.”
“Error catastrophe” is an artifact of the needle-in-a-haystack landscape
Sarkisyan et al. (2016) Local fitness landscape of the green fluorescent protein (*Aequorea victoria*)
Classical multiplicative fitness landscapes defy the lore of the error catastrophe.

Multiplicative Fitness Landscape Model (figurative picture)
Dynamical system combining mutation and natural selection:

\[
\frac{d}{dt} x_i(t) = \sum_{j=1}^{n} M_{ij} w_j x_j(t) - \left( \sum_{j=1}^{n} w_j x_j(t) \right) x_i(t)
\]

or in vector form

\[
\frac{d}{dt} \mathbf{x}(t) = \mathbf{M} \mathbf{D} \mathbf{x}(t) - \mathbf{w}(\mathbf{x}(t)) \mathbf{x}(t)
\]

where

- \( \mathbf{x}(t) \) — vector of genotype frequencies at time \( t \),
- \( \mathbf{M} \) — transmission matrix, \( M_{ij} \) is mutation rate \( j \rightarrow i \),
- \( \mathbf{D} = \text{diag}[w_i] \) — diagonal matrix of fitnesses \( w_i \), and
- \( \mathbf{w}(\mathbf{x}(t)) = \sum_{i=1}^{n} x_i(t) w_i \) — population mean fitness at time \( t \).

**Assumptions:** Infinite population, arbitrary haploid selection, no recombination.
Mutation-selection balance

The population evolves to a stationary distribution, $\hat{x}$, at which $\frac{dx(t)}{dt} = 0$, so the equilibrium $\hat{x}$ satisfies

$$M(\mu)D\hat{x} = \overline{w}(\hat{x}) \hat{x},$$

(1)

hence

- $\hat{x}$ is the *quasispecies* — the Perron vector (dominant eigenvector) of matrix $M(\mu)D$, and

- $\overline{w}(\hat{x}) = r(M(\mu)D)$ is the Perron root (dominant eigenvalue and spectral radius) of $M(\mu)D$.

- **Extinction condition**: Mean fitness is less than one: $\overline{w}(\hat{x}) = r(M(\mu)D) < 1$. (A 5th independent “error threshold”)
Classical multiplicative fitness landscapes defy the lore of the error catastrophe

Multiplicative Fitness Landscape Model (figurative picture)
Lore (Tripathi et al., 2012):

“When the mutation rate is increased beyond a critical value, called the error threshold, the quasispecies delocalizes in sequence space, inducing a severe loss of genetic information—a phenomenon termed error catastrophe—and compromising the viability of the viral population.”
Classical multiplicative fitness landscapes defy the lore of the error catastrophe

Multiplicative fitness counterexample: As the mutation rate increases:

1. genotype and allele frequencies change gradually
2. the mean fitness of the population declines gradually
3. the information content of the population declines gradually
4. no limit is placed on the length of sequences that carry genetic information.
Multiplicative Fitnesses

The diagonal matrix of fitnesses is represented as a Kronecker product,

\[ D = \bigotimes_{\xi=1}^{L} \begin{bmatrix} w & 0 \\ 0 & 1 \end{bmatrix}. \]

The fitness of the binary sequence is

\[ w_i = w^{d_i} \times 1^{L-d_i} = w^{d_i}, \]

where \( d_i \) is the number of 0 alleles in the \( L \)-locus sequence.
Independent Multilocus Mutation

Classical model of mutations occurring independently over \( L \) sites in a genome at mutation rate \( \mu \):

The mutation matrix \( M(\mu) = [M_{ij}(\mu)] \) may be represented using the Kronecker product.

A example with 2-alleles at \( L \) loci:

\[
M(\mu) = \bigotimes_{\xi=1}^{L} \left[ (1 - \mu) \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \mu \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \right]
\]

(2)
The equilibrium relation then becomes

\[ \mathbf{M}(\mu)\mathbf{D}\hat{x} = \bigotimes_{\xi=1}^{L} \left[ (1 - \mu) \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \mu \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \right] \bigotimes_{\xi=1}^{L} \begin{bmatrix} w & 0 \\ 0 & 1 \end{bmatrix} \hat{x} \]

\[ = \bigotimes_{\xi=1}^{L} \left\{ (1 - \mu) \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \mu \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \right\} \begin{bmatrix} w & 0 \\ 0 & 1 \end{bmatrix} \hat{x} \]

\[ = r(\mathbf{M}(\mu)\mathbf{D}) \hat{x}, \]

where \( r(\mathbf{M}(\mu)\mathbf{D}) \) is the spectral radius of \( \mathbf{M}(\mu)\mathbf{D} \) — the asymptotic aggregate growth rate of the quasispecies.
Because of the Kronecker product form, the equilibrium distribution $\hat{\mathbf{x}}$ also factors into

$$
\hat{\mathbf{x}} = \bigotimes_{\xi=1}^L \hat{\mathbf{g}} = \bigotimes_{\xi=1}^L \begin{bmatrix} \hat{g}_0 \\ \hat{g}_1 \end{bmatrix},
$$

where $\hat{g}_0 = 1 - \hat{g}_1$, solves the single-locus equilibrium relation

$$
\left( (1-\mu) \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \mu \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \right) \begin{bmatrix} w & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{g}_0 \\ \hat{g}_1 \end{bmatrix} = \overline{w}(\hat{\mathbf{g}}) \begin{bmatrix} \hat{g}_0 \\ \hat{g}_1 \end{bmatrix},
$$

with $\overline{w}(\hat{\mathbf{g}}) = w\hat{g}_0 + \hat{g}_1$. 
The closed form solution (Woodcock and Higgs, 1996) is

$$\hat{g} = \begin{bmatrix} \hat{g}_0 \\ \hat{g}_1 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} + \frac{\mu (w + 1) - c}{2(1 - w)} \\ \frac{1}{2} - \frac{\mu (w + 1) - c}{2(1 - w)} \end{bmatrix},$$

where $c := \sqrt{(1 - \mu)^2 (w + 1)^2 - 4w(1 - 2\mu)}$. 
Single-locus equilibrium frequency of allele 1 plotted as a function of mutation rate $\mu$ and selection coefficient $w$.
Quantifying Genetic Information in a Population

- The Kullback-Leibler divergence between stationary distributions with and without natural selection acting (Strelioff et al., 2010; Schuster, 2013):

\[ I(\hat{x}) := D_{KL}(\hat{x} \parallel \pi) = D_{KL}(\hat{x} \parallel 2^{-L} \mathbf{e}) \]

\[ = \sum_{i=1}^{2^L} \hat{x}_i \log_2 \frac{\hat{x}_i}{2^{-L}} = L + \sum_{i=1}^{2^L} \hat{x}_i \log_2 \hat{x}_i \]

\[ = L - \mathcal{H}(\hat{x}), \]

where

- \( \pi \) is the stationary distribution without selection
- \( \hat{x} \) is the stationary distribution with selection
- \( \mathcal{H}(\hat{x}) \) is the Shannon entropy of \( \hat{x} \), and \( \mathbf{e} \) is the vector of ones.
Genetic information at mutation-selection balance, $\hat{x}$

- Equilibrium frequencies: $\hat{x}_i = \hat{g}_0^{d_i} \hat{g}_1^{L-d_i} = (1 - \hat{g}_1)^{d_i} \hat{g}_1^{L-d_i}$.

- Information generated by natural selection:

\[
\mathcal{I}(\hat{x}) = L + \sum_{i=1}^{2^L} (1 - \hat{g}_1)^{d_i} \hat{g}_1^{L-d_i} [d_i \log_2(1 - \hat{g}_1) + (L - d_i) \log_2 \hat{g}_1]
\]

\[
= L \mathcal{I}(\hat{g}).
\]

Genetic information in the population is simply the genetic information at each locus times the number of loci $L$. 
Single-locus **genetic information** as a function of mutation rate $\mu$ and selection coefficient $w$. 

$$I(\hat{g})$$

$\mu$  MUTATION RATE

$w$  FITNESS
Genetic information maintained in a genome of length 10,000 bases, mutation rate $= 0.1$ per base

Number of bits maintained by selection as a function of per-base selection coefficient $w$.

\[ I(\hat{x}) = 0.1 \]
Genetic information maintained in a genome of length 10,000 bases, mutation rate $= 0.01$ per base

Number of bits maintained by selection as a function of per-base selection coefficient $w$. 

![Graph showing the relationship between $\mathcal{I}(\hat{x})$, $w$, and fitness with $\mu = 0.01$.]
Example 2: A Quasispecies “Yo-Yo”

- Genotype fitness is a function of the number of mutations to the “master sequence” 000000000000000.

- 15 loci, 2 alleles per locus, i.i.d. mutation at rate $\mu$ per locus.
Result: Multiple, Reversing “Error Catastrophes” — A “Yo-Yo”

**STATIONARY DISTRIBUTION vs. MUTATION RATE**

**SEQUENCES WITH:**
- MAJORITY 0 ALLELES
- MAJORITY 1 ALLELES

- FREQUENCY
- MUTATION RATE \( \mu \rightarrow \)

- Stationary distribution vs. mutation rate
- Sequences with:
  - Majority 0 alleles
  - Majority 1 alleles

- The graph shows the relationship between the frequency of sequences and mutation rate, demonstrating how the distribution changes with varying mutation rates.
Result: Multiple, Reversing “Error Catastrophes” — A “Yo-Yo”

<table>
<thead>
<tr>
<th>MUTATION RATE $\mu$</th>
<th>AVERAGE DENSITY OF 1s</th>
<th>NUMBER OF 1s</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>0.00</td>
<td>000000000000000</td>
</tr>
<tr>
<td>0.02</td>
<td>0.04</td>
<td>111111111111111</td>
</tr>
<tr>
<td>0.04</td>
<td>0.06</td>
<td>0.02 0.04 0.06 0.08</td>
</tr>
<tr>
<td>0.06</td>
<td>0.08</td>
<td>0.2 0.4 0.6 0.8 1.0</td>
</tr>
<tr>
<td>0.08</td>
<td>0.10</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Lee Altenberg | How Much Information Can Natural Selection Maintain? | 37/42
Fitness Declines Smoothly Despite Four Error Thresholds

- Drama in the genotype frequencies contrasts with Smooth decline of the quasispecies fitness $r(M(\mu)D)$
Sequence information in the population declines (‘delocalizes’) gradually with mutation rate, except at the error thresholds where it dips but bounces back again with increasing mutation.
Conclusion 1

It is not in general true that:

1. there is a critical mutation rate, an “error threshold” above which all genetic information in a population is lost — the “error catastrophe”

2. the mutation rate restricts the length of sequences that can be replicated.

Thus, claims in the literature that viruses “replicate near the error threshold” may not even be defined.
Instead, it is possible (shown by the multiplicative landscape and other examples (Schuster, 2013)) that:

1. the genetic information in a population degrades gradually as a function of mutation rate, and
2. even at very high mutation rates, long sequences may be reproduced which have low genetic information density,
3. but which have high total information content.

This is just an illustrative example. Characterizing the information dynamics of different fitness landscapes remains an unexplored open question.

I gratefully acknowledge support from:

- Office of the Vice Chancellor for Research, University of Hawai‘i at Mānoa
- Stanford Center for Computational, Evolutionary, and Human Genomics
- Konrad Lorenz Institute for Evolution and Cognition Research, Klosterneuburg, Austria
- Mathematical Biosciences Institute, Columbus, Ohio

Thank you for your attention!
References I


MORE QUOTES
Arias et al. (2013) Molecular dissection of a viral quasispecies under mutagenic treatment: positive correlation between fitness loss and mutational load:

“Theoretical predictions suggested that these error frequencies in RNA viruses are near to a maximum value compatible with maintaining genetic information and therefore, virus viability, namely the error threshold.”
“excessive mutation can abrogate hill climbing, replacing selection with diffusion over the simplex.

This is known as the ‘error threshold.’

For any choice of fitness function, the regime $p > 1/L$ will completely ‘flatten’ the landscape, eliminating adaptation altogether.”