Mānoa Mini-Symposium on Physics of Adaptive Computation

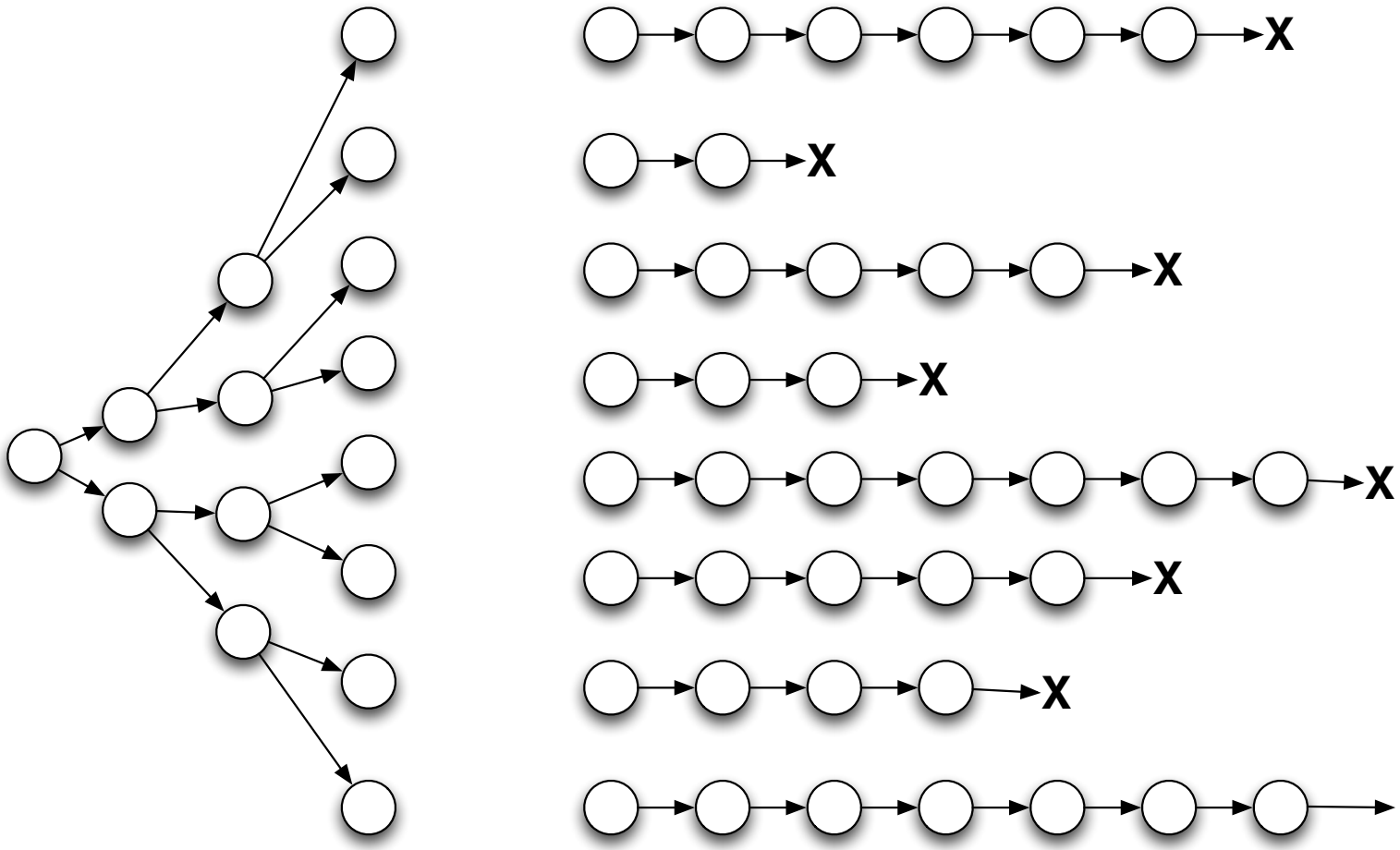
How Much Information Can Natural Selection Maintain?

January 7, 2019

Lee Altenberg
altenber@hawaii.edu
http://dynamics.org/Altenberg/

Fundamental Processes in Biology: (1) Growth

GROWTH **DECAY (NEGATIVE GROWTH)**

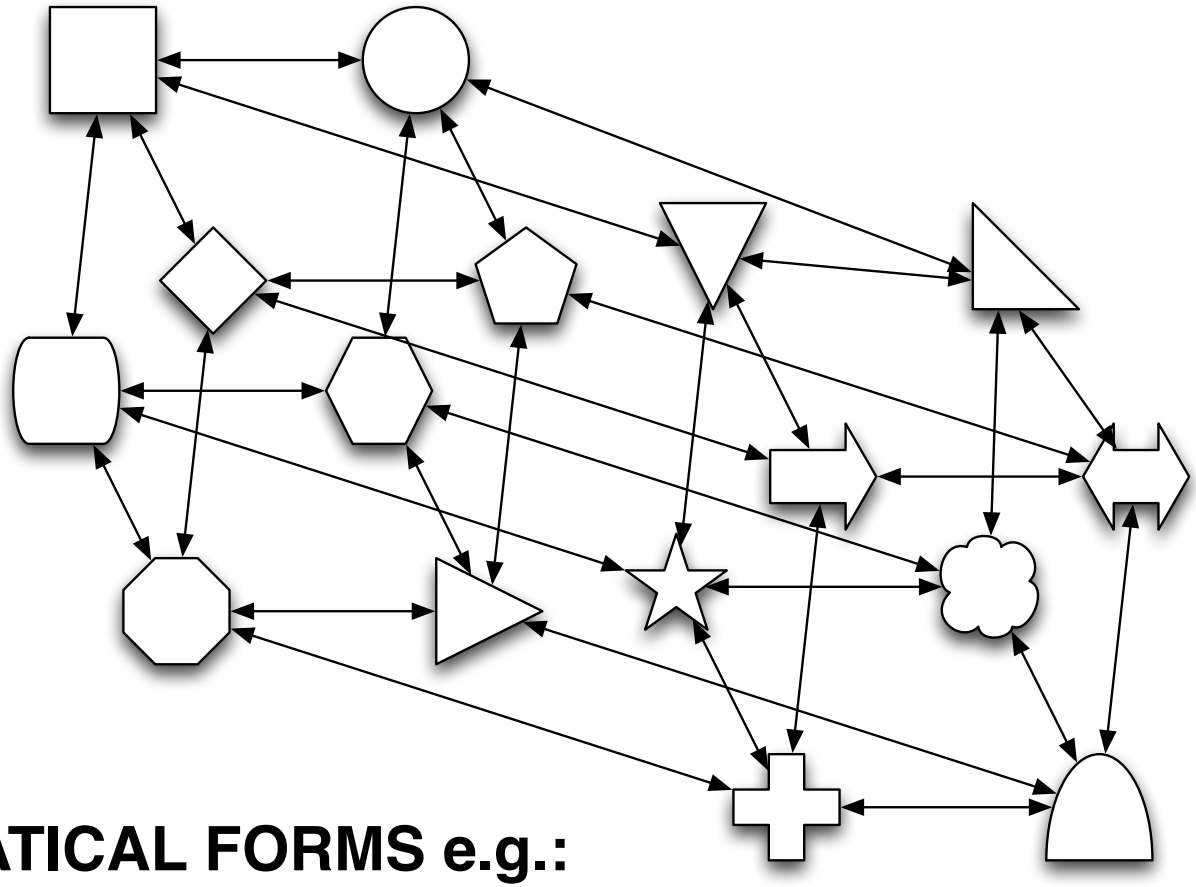


MATHEMATICAL FORMS e.g.:

$$x(t+1) = \lambda\, x(t) \qquad \frac{\mathrm{d}}{\mathrm{d}t}x(t) = \lambda\, x(t)$$

Fundamental Processes in Biology: (2) Transformation

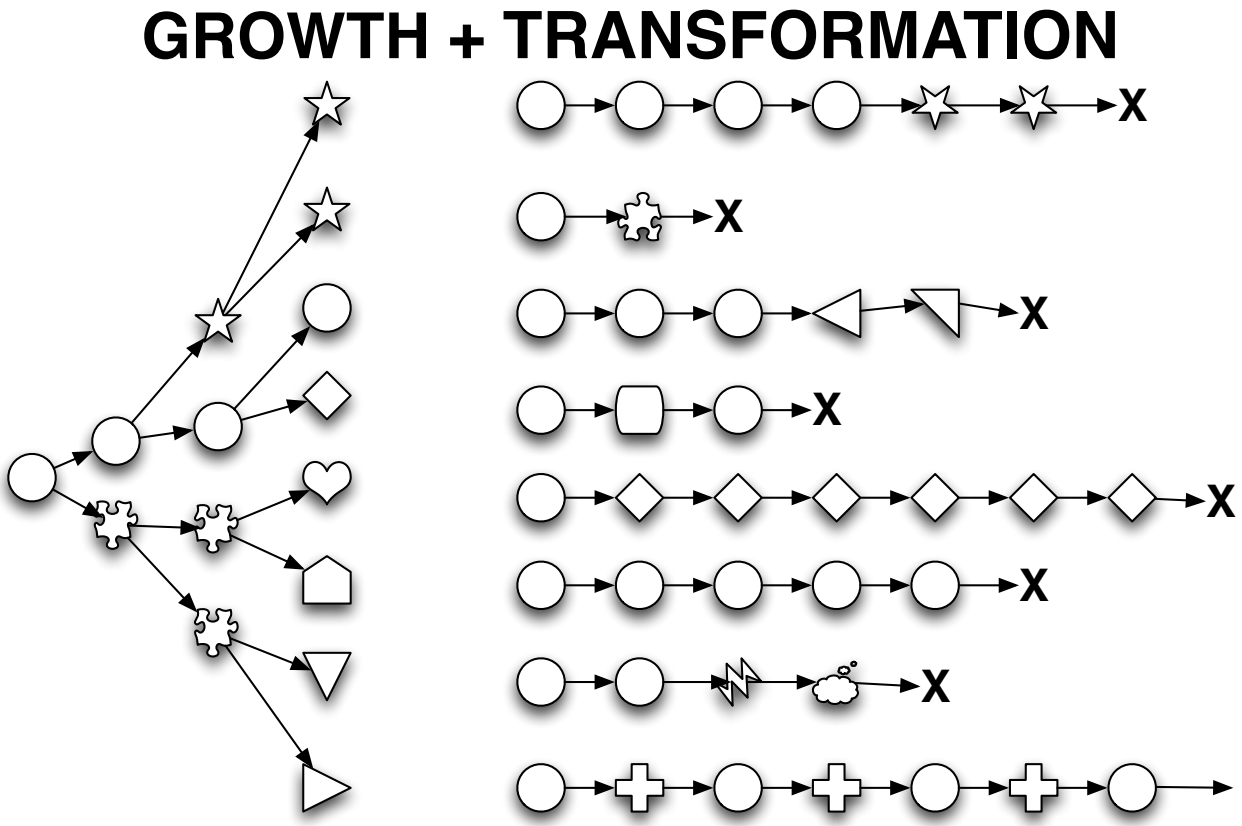
TRANSFORMATION (i.e. CHANGE OF STATE)



MATHEMATICAL FORMS e.g.:

$$x(t+1) = \sum_y P_{xy}\ y(t) \qquad \frac{\partial}{\partial t} u(x,t) = \Delta u(x,t)$$

Growth and Transformation Combined

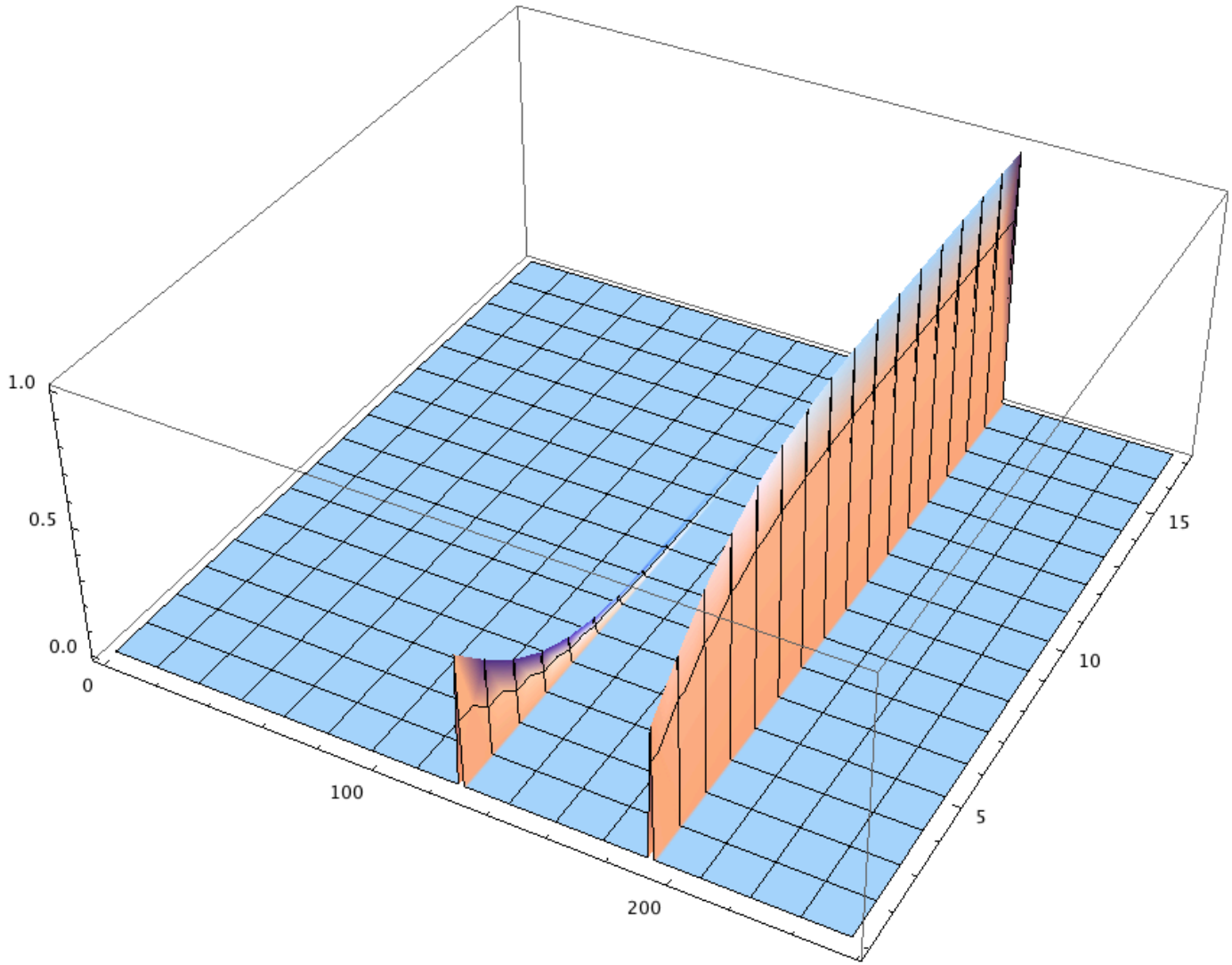


MATHEMATICAL FORMS e.g.:

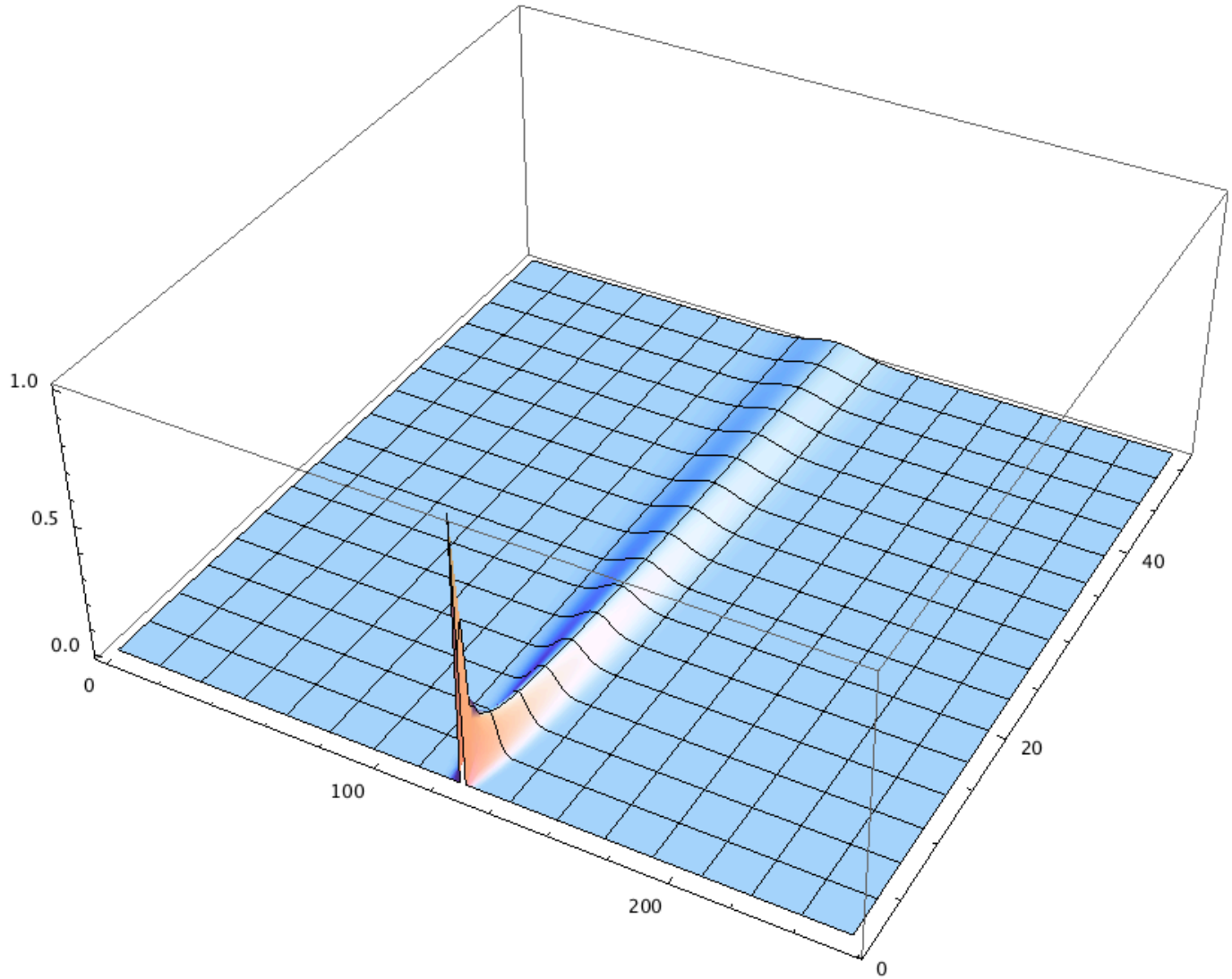
$$x(t+1) = \sum_y P_{xy}\ \lambda_y\ y(t)$$

$$\frac{\partial}{\partial t} u(x,t) = \Delta_x\ u(x,t) + \lambda(u)\ u(x,t)$$

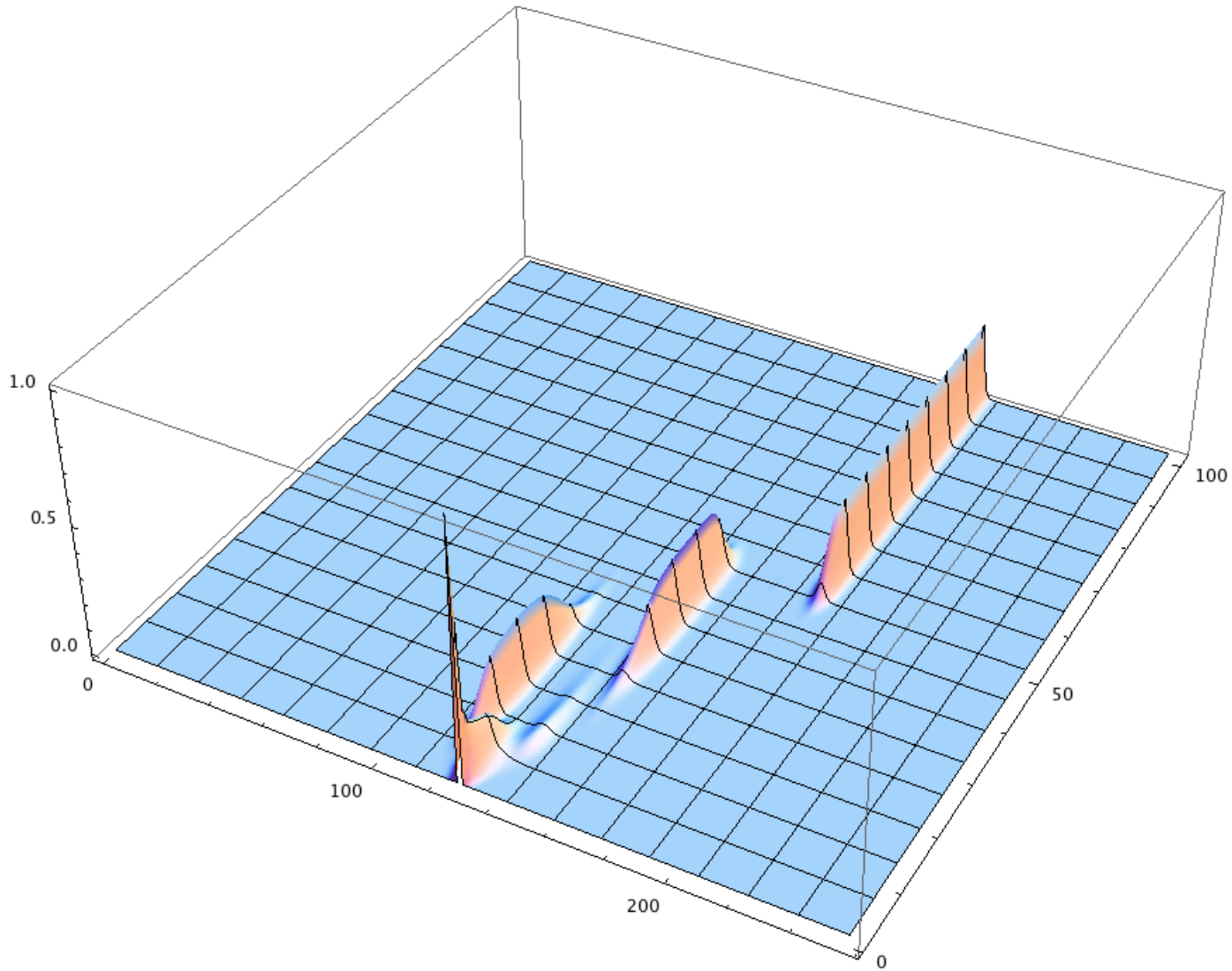
Growth is a *concentrating* operator



Transformation is a *dispersing* operator



Growth and Transformation combined create a *search* operator



Basic question regarding the information stored by organisms:

Eigen and Schuster (1977) in "The Hypercycle: A principle of natural self-organization."

What is the relationship between

1. mutation,
2. natural selection, and
3. the accumulation of information in the genome?

The Quasispecies, Eigen and Schuster (1977)

- "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.

The Quasispecies, Eigen and Schuster (1977)

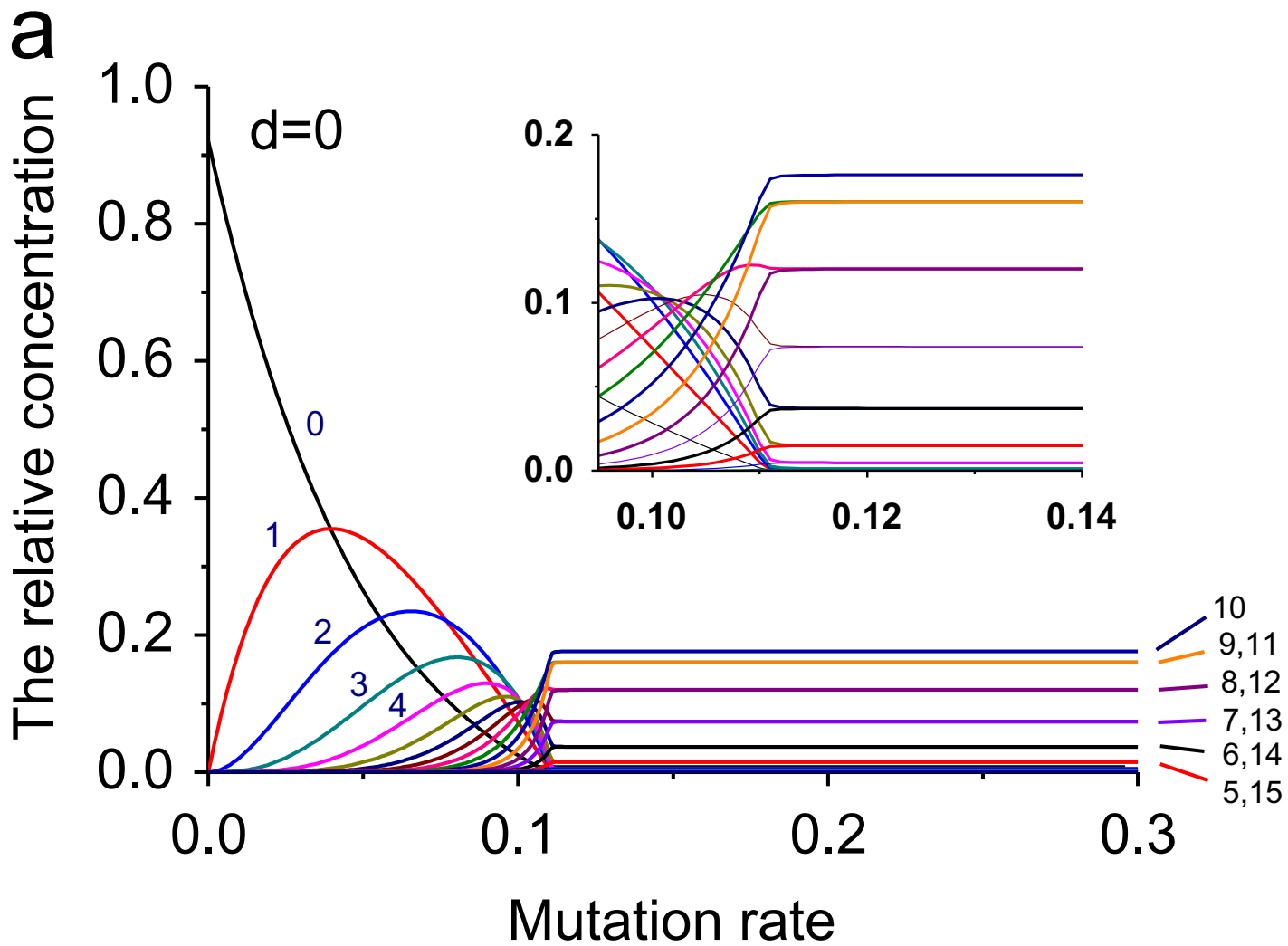
- "If these criteria are violated, the information stored in the nucleotide sequence of the master copy will disintegrate irreversibly leading to an error catastrophe.
- As a consequence, selection and evolution of RNA or DNA molecules is limited with respect to the amount of information that can be stored in a single replicative unit."

Limits on the Length of Replicating Sequences

Eigen and Schuster (1977, p. 555):

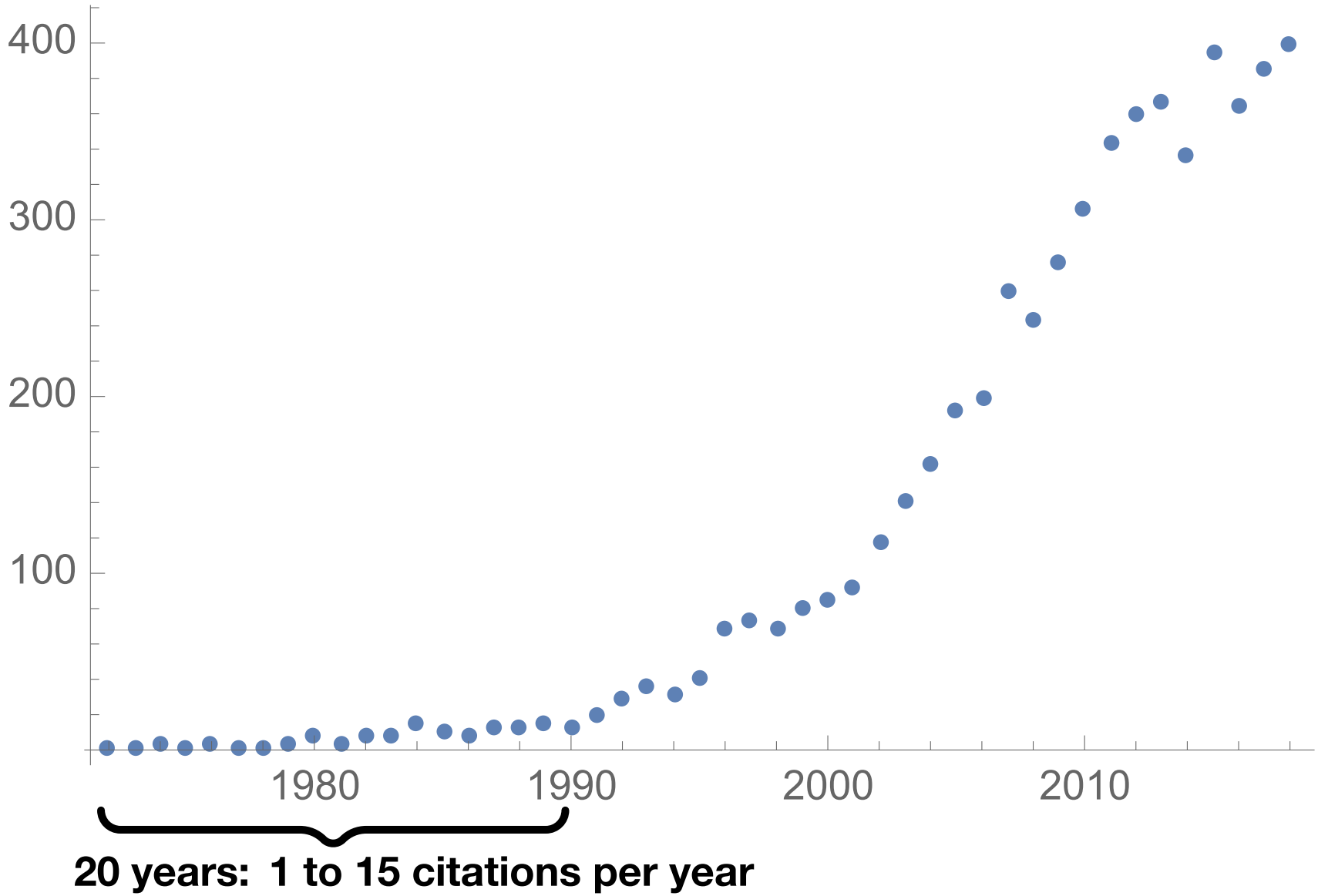
- "There is a threshold-relationship for the rate of mutation, at which evolution is fastest, but which must not be surpassed unless all the information thus far accumulated in the evolutionary process is to be lost."
- "The number of molecular symbols of a self-reproducible unit is restricted, the limit 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



Papers per year with "error catastrophe" or "error threshold"

AND (quasispecies OR Eigen)



Hermisson et al. (2002) Mutation–Selection Balance: Ancestry, Load, and Maximum Principle

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], and
4. a jump in the population mean of the mutational distance"

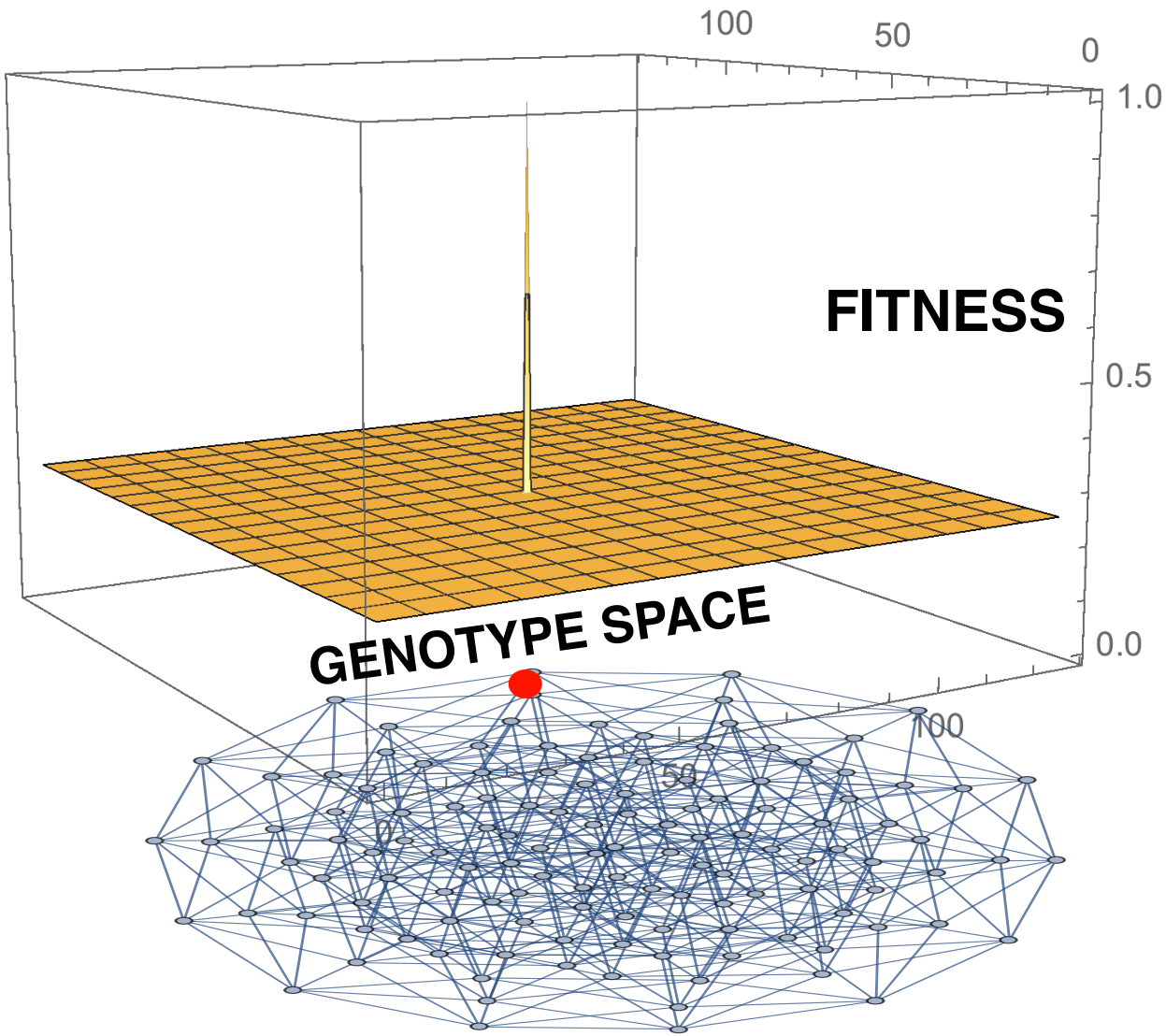
In Eigen and Schuster's fitness landscape, these four "error thresholds" happen to coincide. But they may not even exist in the general case.

The Lore: These insights and caveats still have not penetrated the literature

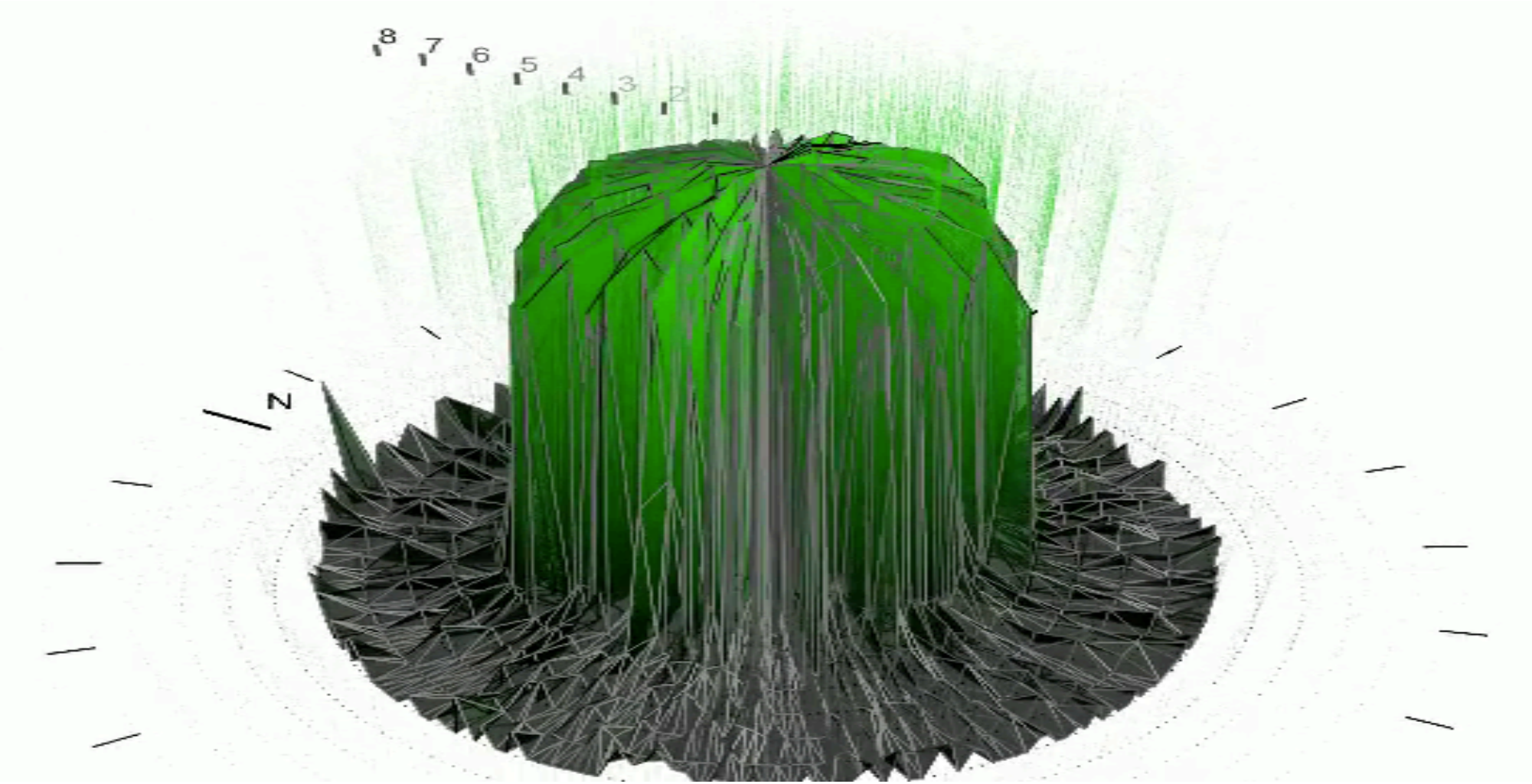
Ten years after Hermisson et al. (2002) we still find "the lore":
e.g. Barbieri (2012) Code biology–A new science of life

> "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

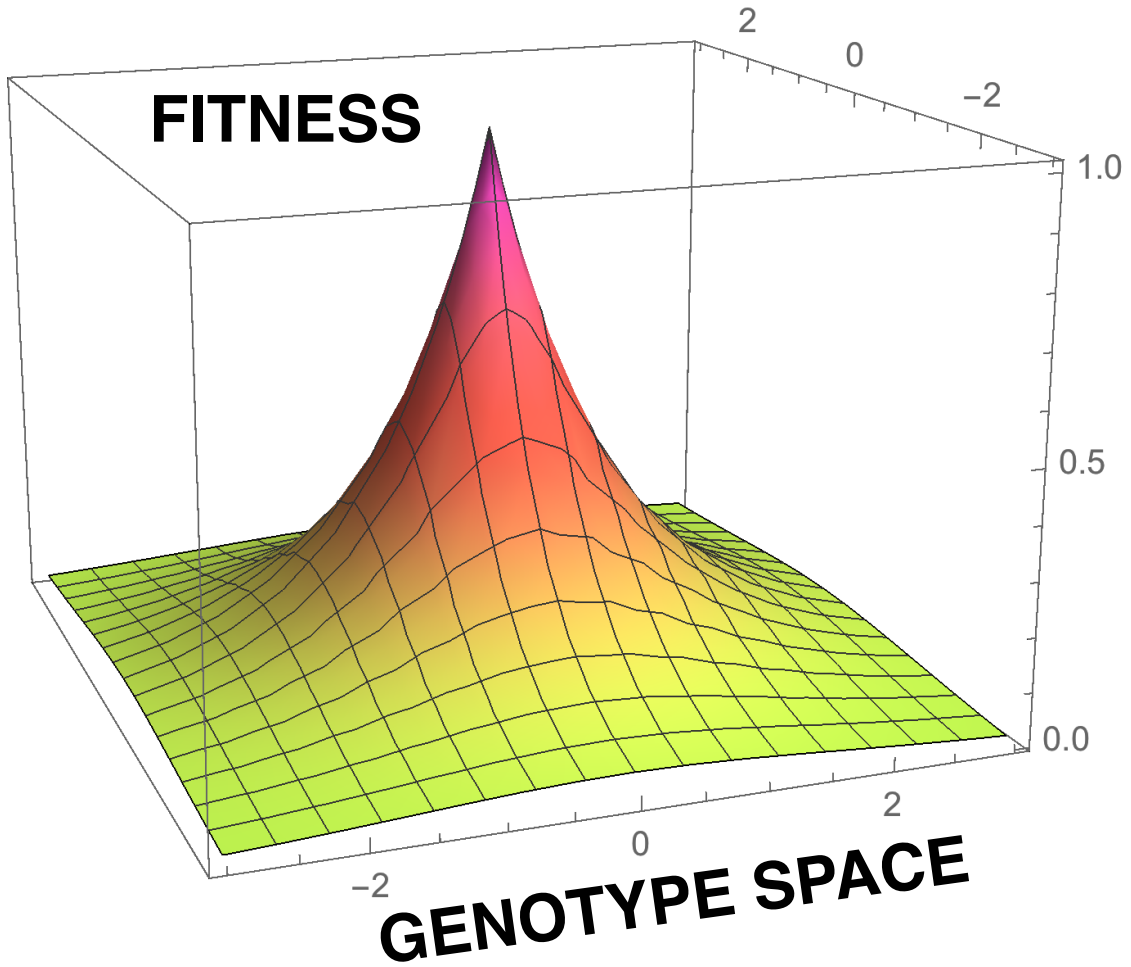


Sarkisyan et al. (2016) Local fitness landscape of the green fluorescent protein (*Aequorea victoria*)



Classical multiplicative fitness landscapes defy the lore of the error catastrophe

Multiplicative Fitness Landscape Model (figurative picture)



Dynamical system combining mutation and natural selection:

$$\frac{\mathrm{d}}{\mathrm{d}t}x_i(t) = \sum_{j=1}^{n} M_{ij} w_j x_j(t) - \left(\sum_{j=1}^{n} w_j x_j(t) \right) \; x_i(t)$$

or in vector form

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}(t) = \mathbf{M}\,\mathbf{D}\,\mathbf{x}(t) - \overline{w}(\mathbf{x}(t))\;\mathbf{x}(t)$$

where

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

Assumptions: Infinite population, arbitrary haploid selection, no recombination.

Mutation-selection balance

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

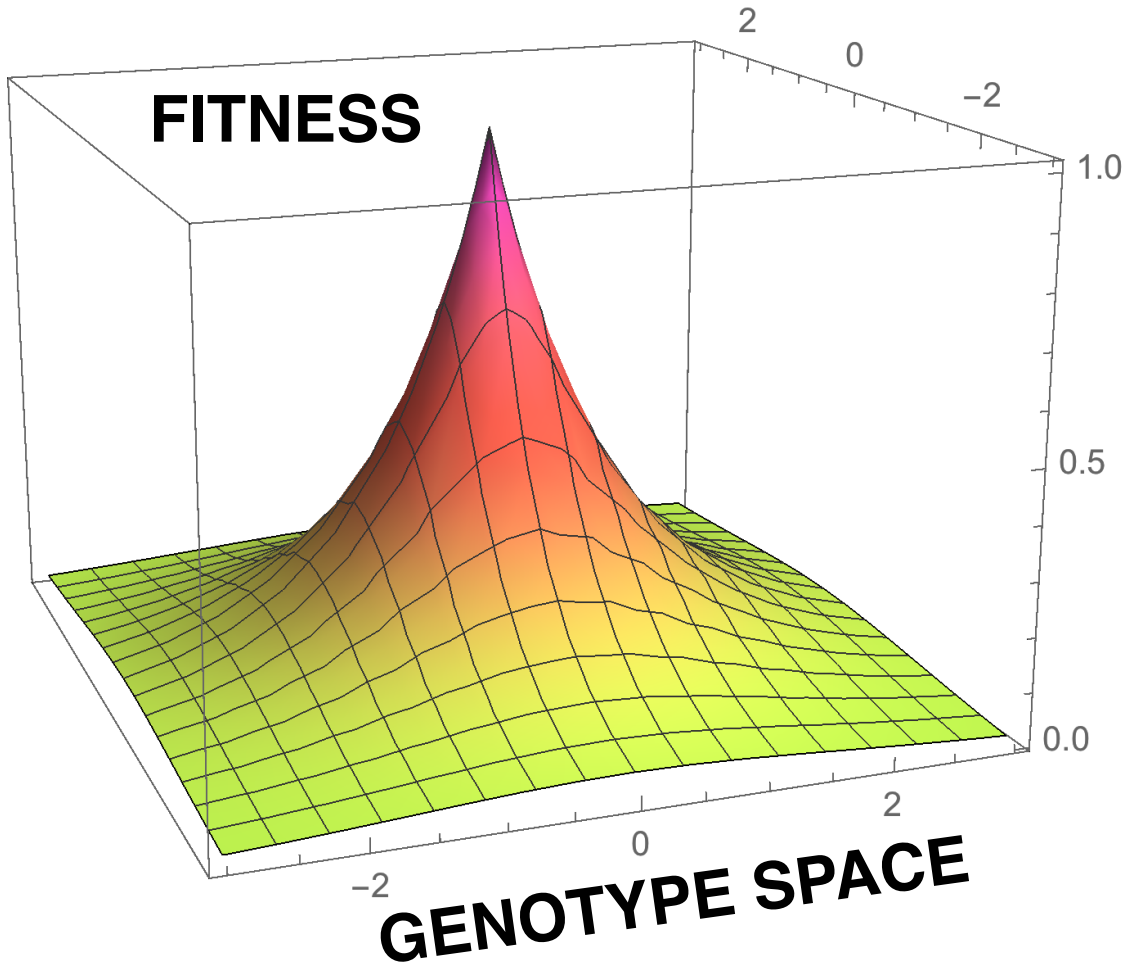
$$\mathbf{M}(\mu)\mathbf{D}\hat{\mathbf{x}} = \overline{w}(\hat{\mathbf{x}})\ \hat{\mathbf{x}}, \tag{1}$$

hence

- $\hat{\mathbf{x}}$ is the *quasispecies* — the Perron vector (dominant eigenvector) of matrix $\mathbf{M}(\mu)\mathbf{D}$, and
- $\overline{w}(\hat{\mathbf{x}}) = r(\mathbf{M}(\mu)\mathbf{D})$ is the Perron root (dominant eigenvalue and spectral radius) of $\mathbf{M}(\mu)\mathbf{D}$.
- Extinction condition: Mean fitness is less than one: $\overline{w}(\hat{\mathbf{x}}) = r(\mathbf{M}(\mu)\mathbf{D}) < 1$. (A 5th independent "error threshold")

Classical multiplicative fitness landscapes defy the lore of the error catastrophe

Multiplicative Fitness Landscape Model (figurative picture)



Reprise: The lore of the error catastrophe

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,

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

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 μ:
The mutation matrix $\mathbf{M}(\mu) = [M_{ij}(\mu)]$ may be represented using the Kronecker product.
A example with 2-alleles at L loci:

$$\mathbf{M}(\mu) = \bigotimes_{\xi=1}^{L} \left[(1-\mu) \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \mu \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \right] \tag{2}$$

The equilibrium relation then becomes

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

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

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

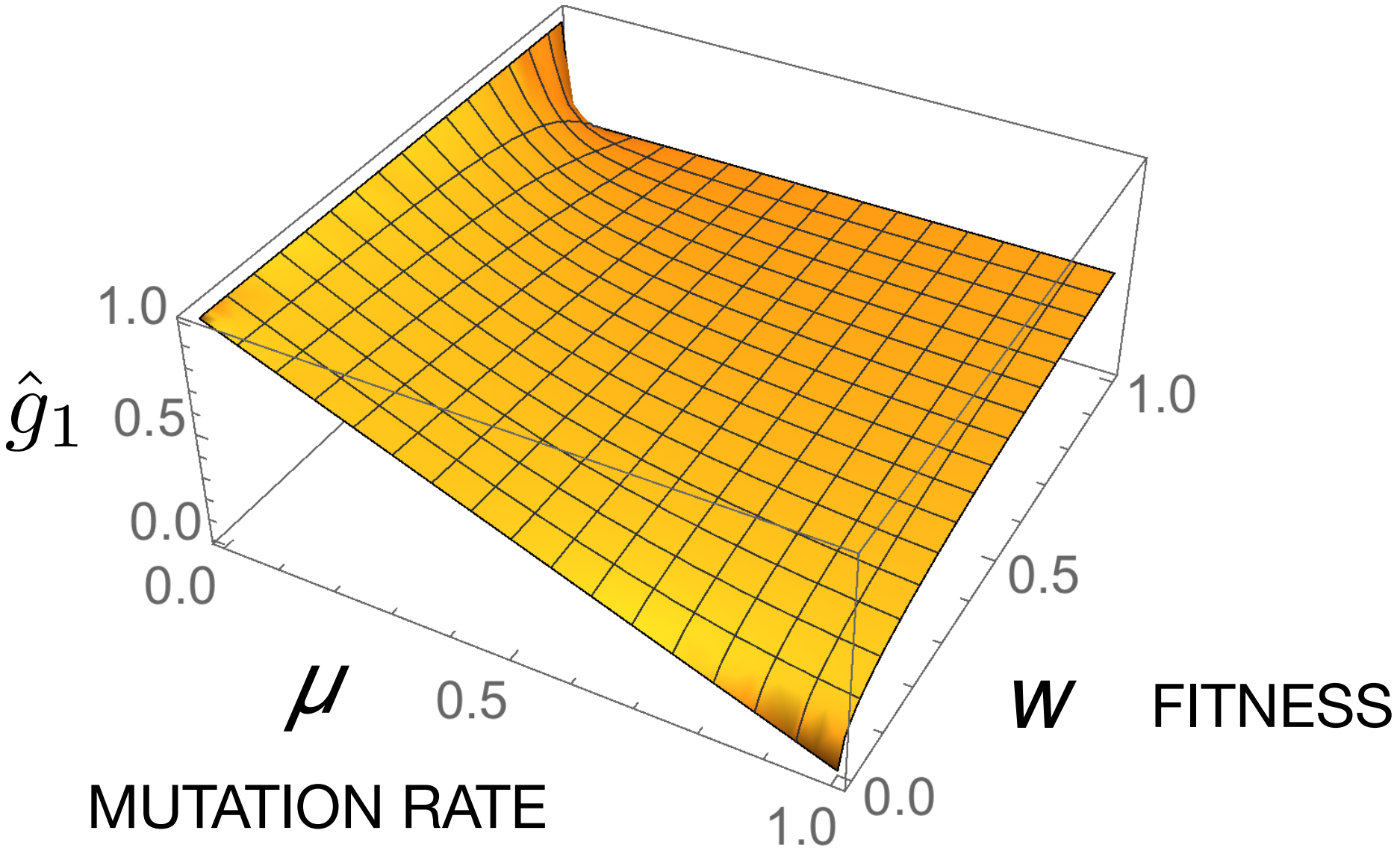
with $\overline{w}(\hat{\mathbf{g}}) = w\hat{g}_0 + \hat{g}_1$.

The closed form solution (Woodcock and Higgs, 1996) is

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

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 μ 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):

$$\begin{aligned}\mathcal{I}(\hat{\mathbf{x}}) &:= D_{\mathrm{KL}}(\hat{\mathbf{x}} \,||\, \boldsymbol{\pi}) = D_{KL}(\hat{\mathbf{x}} \,||\, 2^{-L}\,\mathbf{e}) \\ &= \sum_{i=1}^{2^L} \hat{x}_i \log_2 \frac{\hat{x}_i}{2^{-L}} = L + \sum_{i=1}^{2^L} \hat{x}_i \log_2 \hat{x}_i \\ &= L - \mathcal{H}(\hat{\mathbf{x}}),\end{aligned}$$

where

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

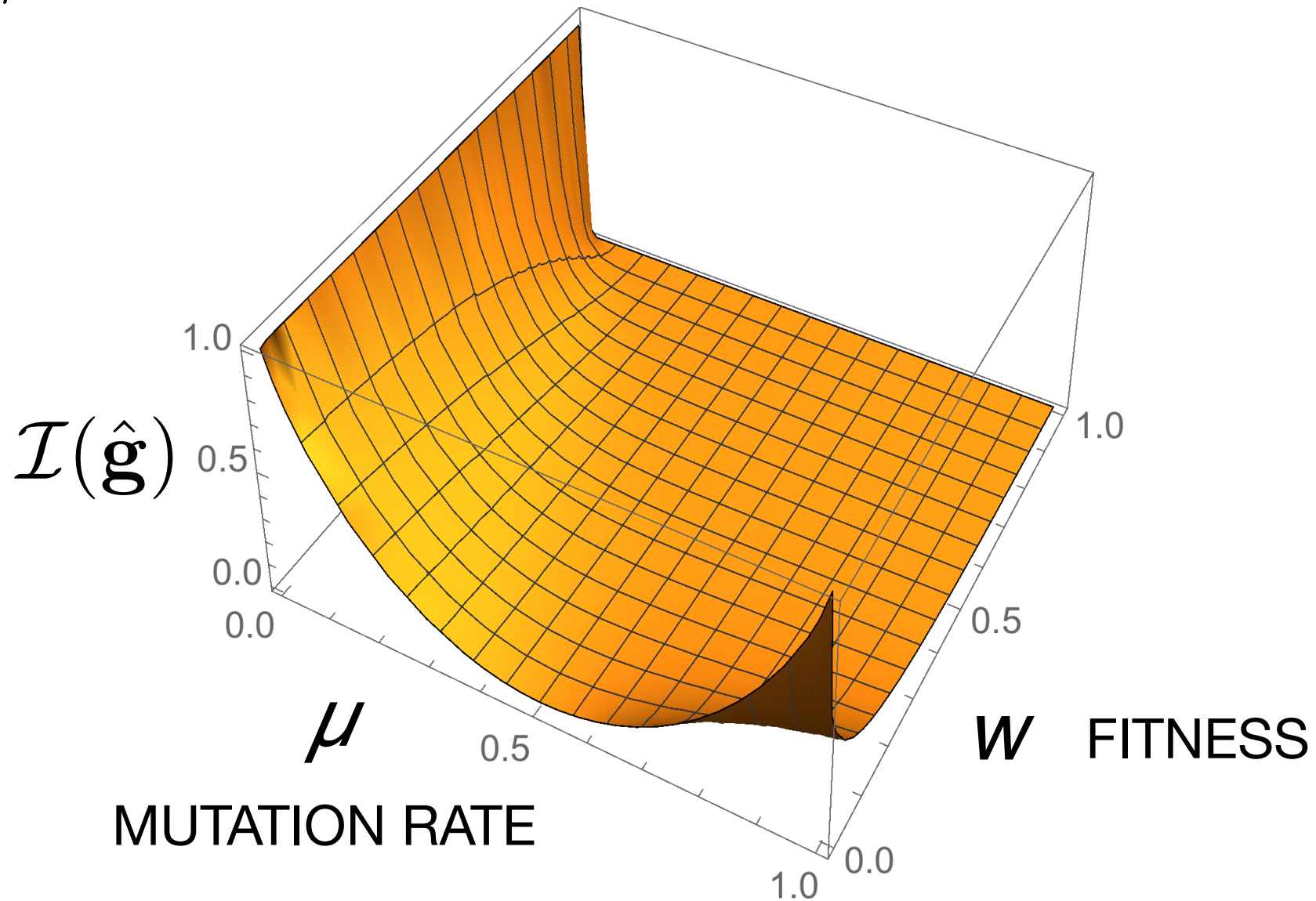
Genetic information at mutation-selection balance, $\hat{\mathbf{x}}$

- 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:

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

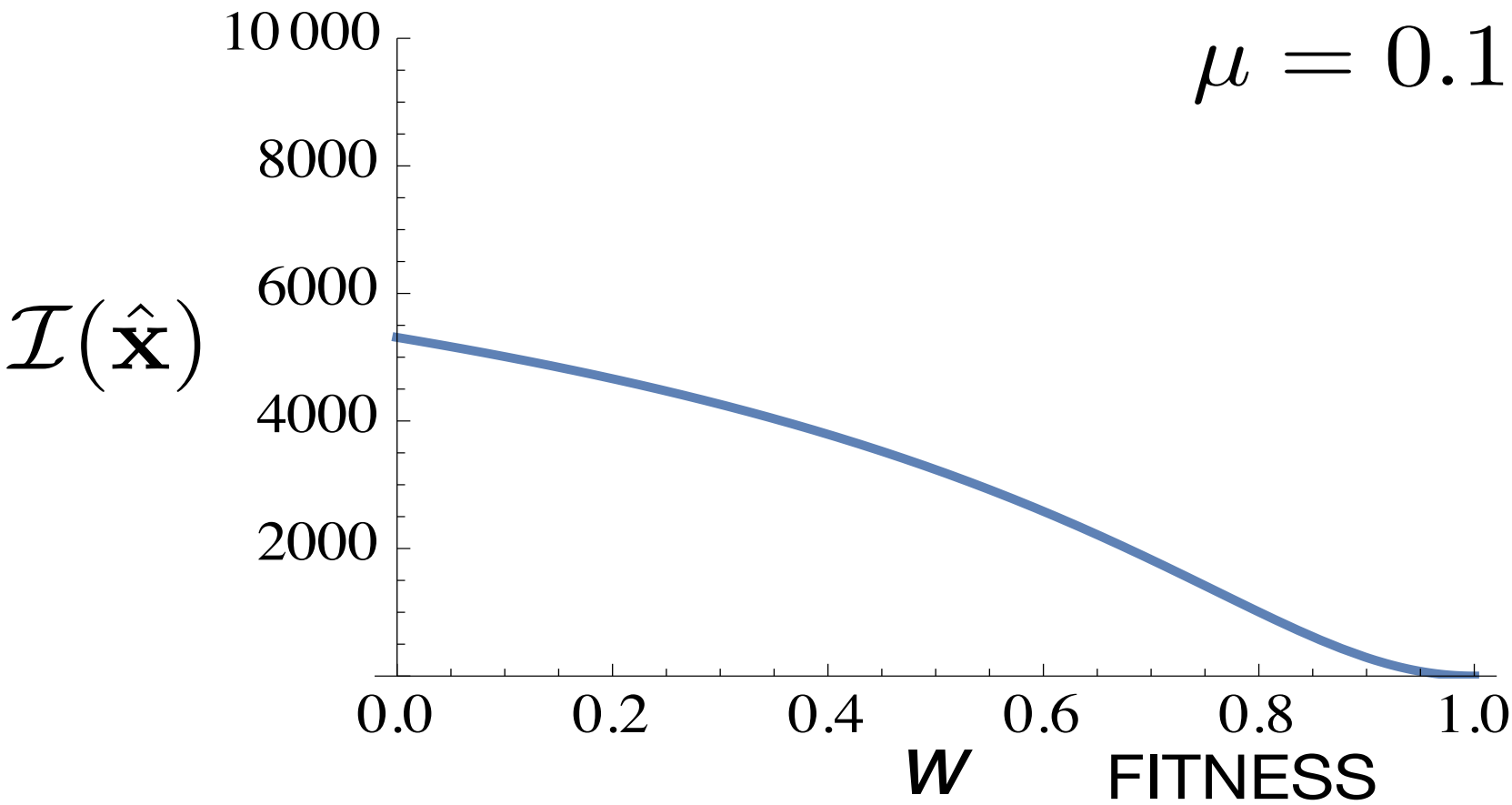
Genetic information in the population is simply the genetic information at each locus times the number of loci L.

Single-locus genetic information as a function of mutation rate μ and selection coefficient w.



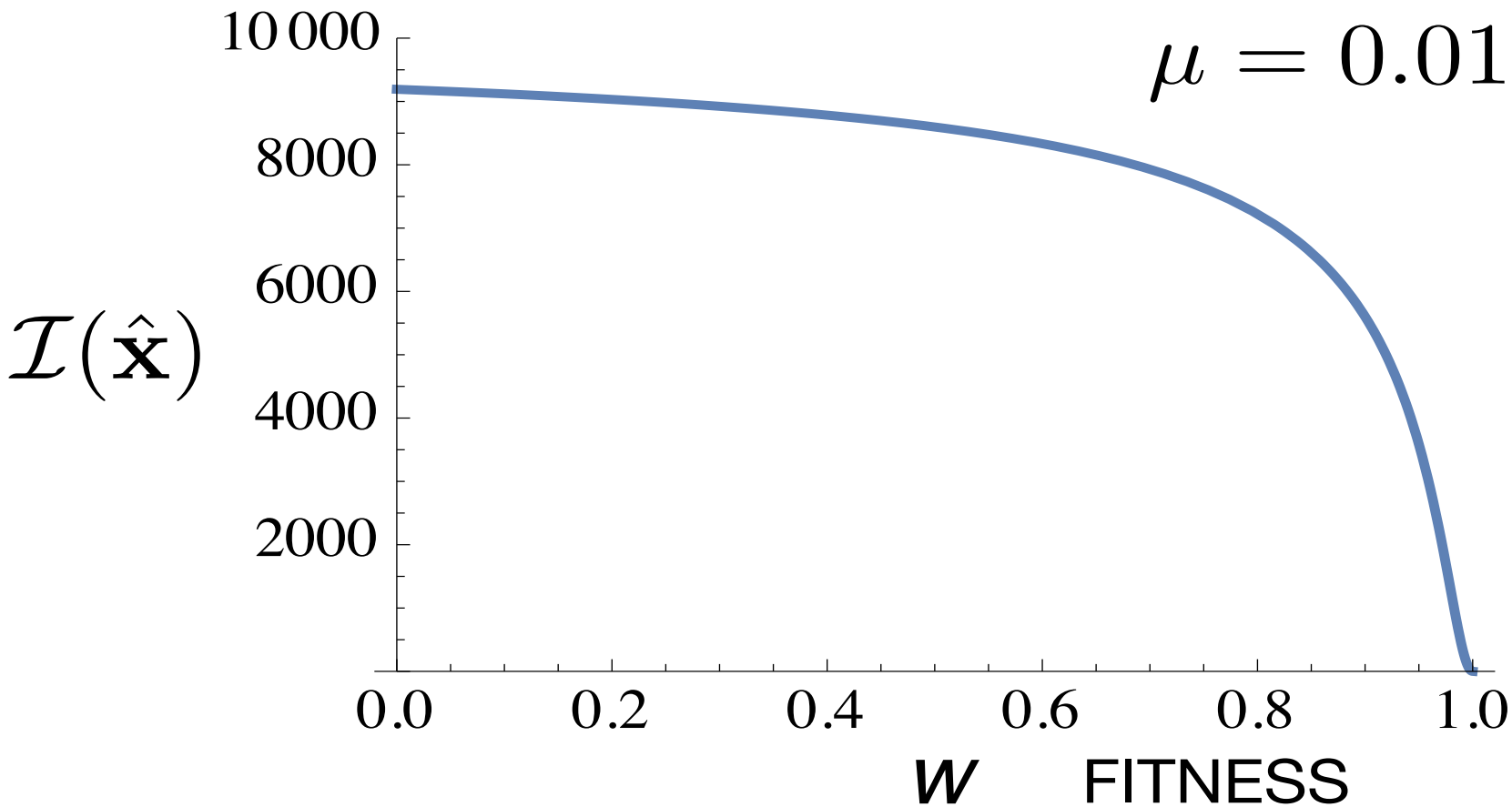
Genetic information maintained in a genome of length 10,000 bases, mutation rate= 0.1 per base

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



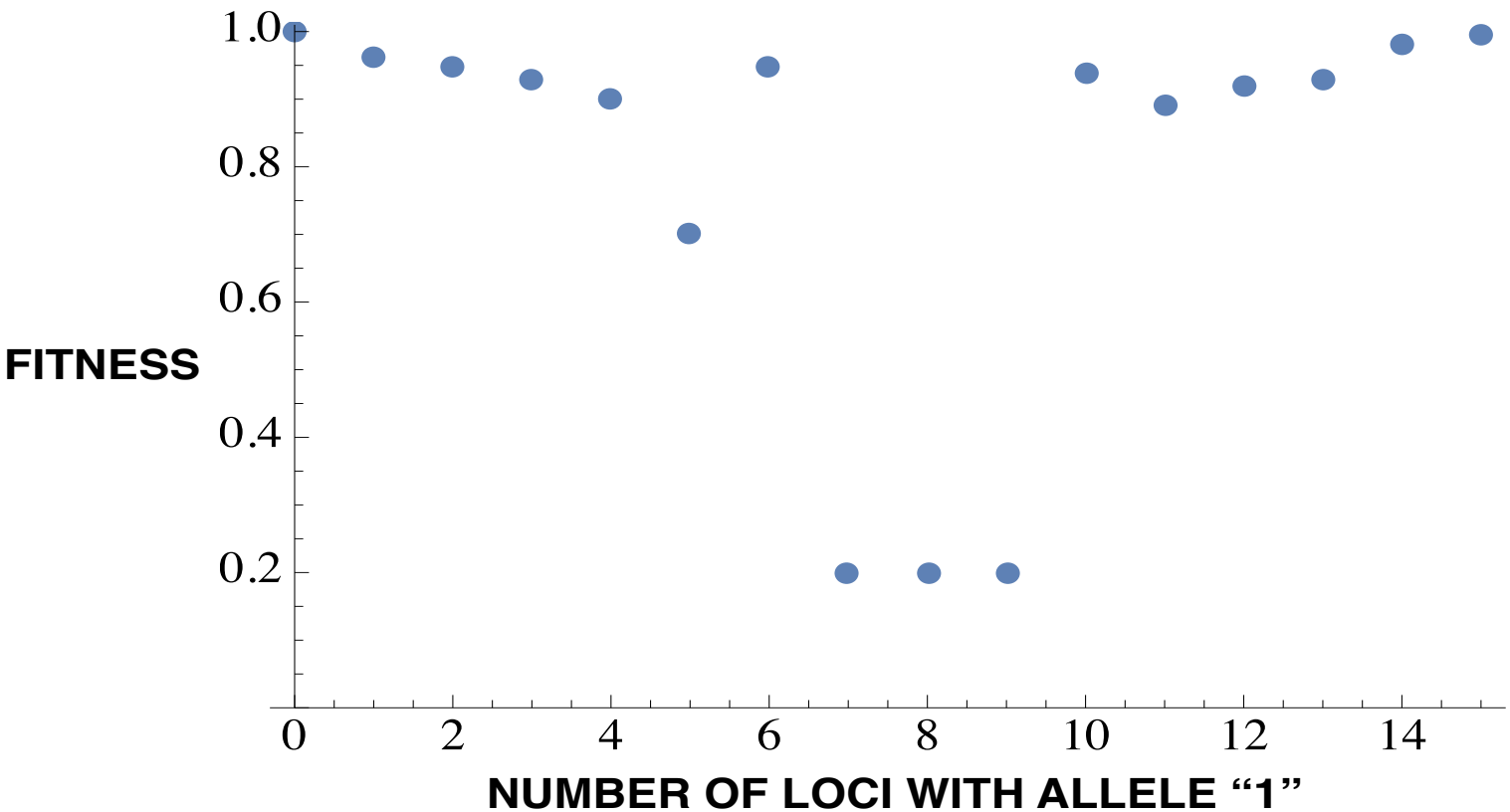
Genetic information maintained in a genome of length 10,000 bases, mutation rate$= 0.01$ per base

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

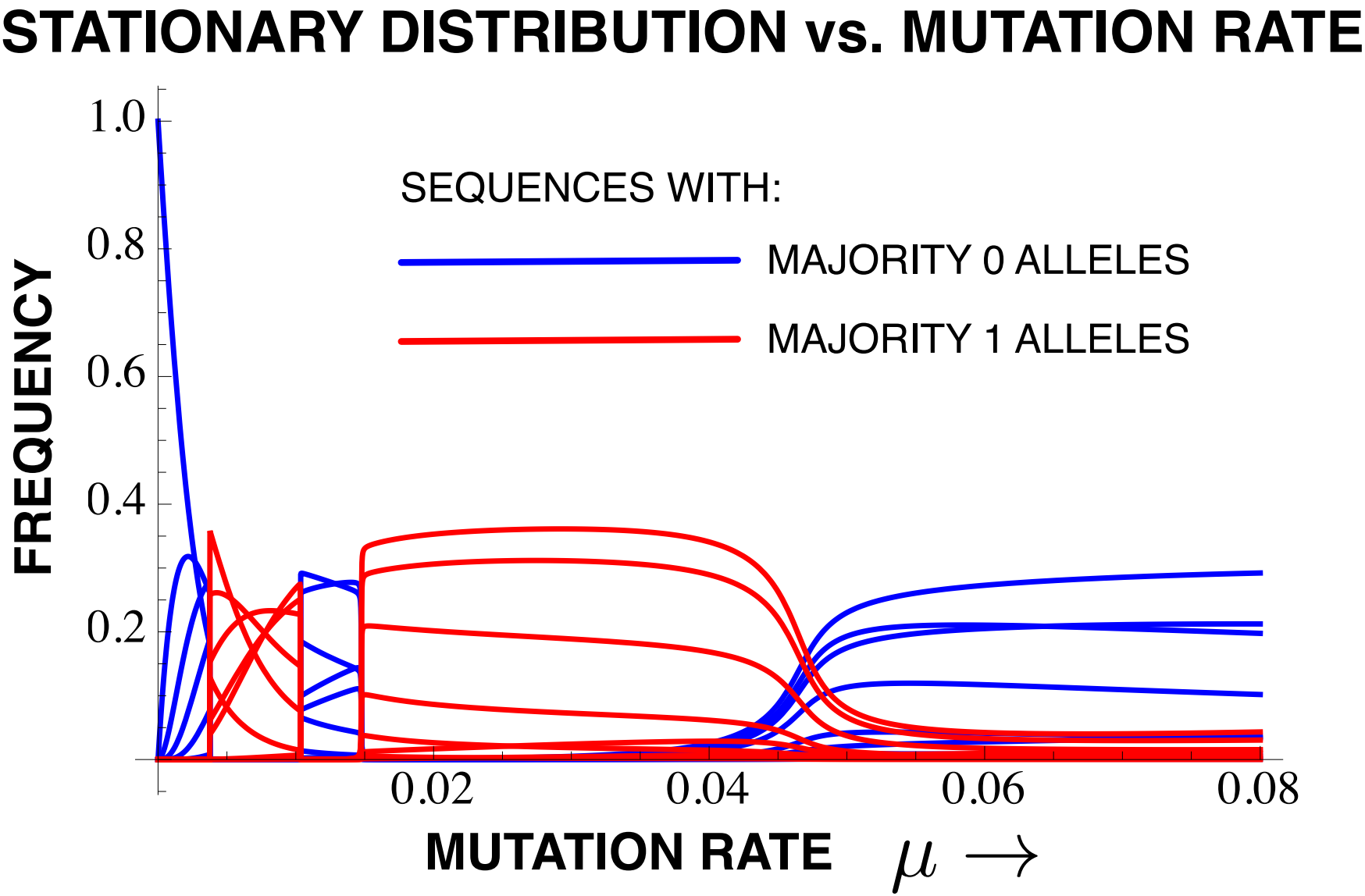


Example 2: A Quasispecies "Yo-Yo"

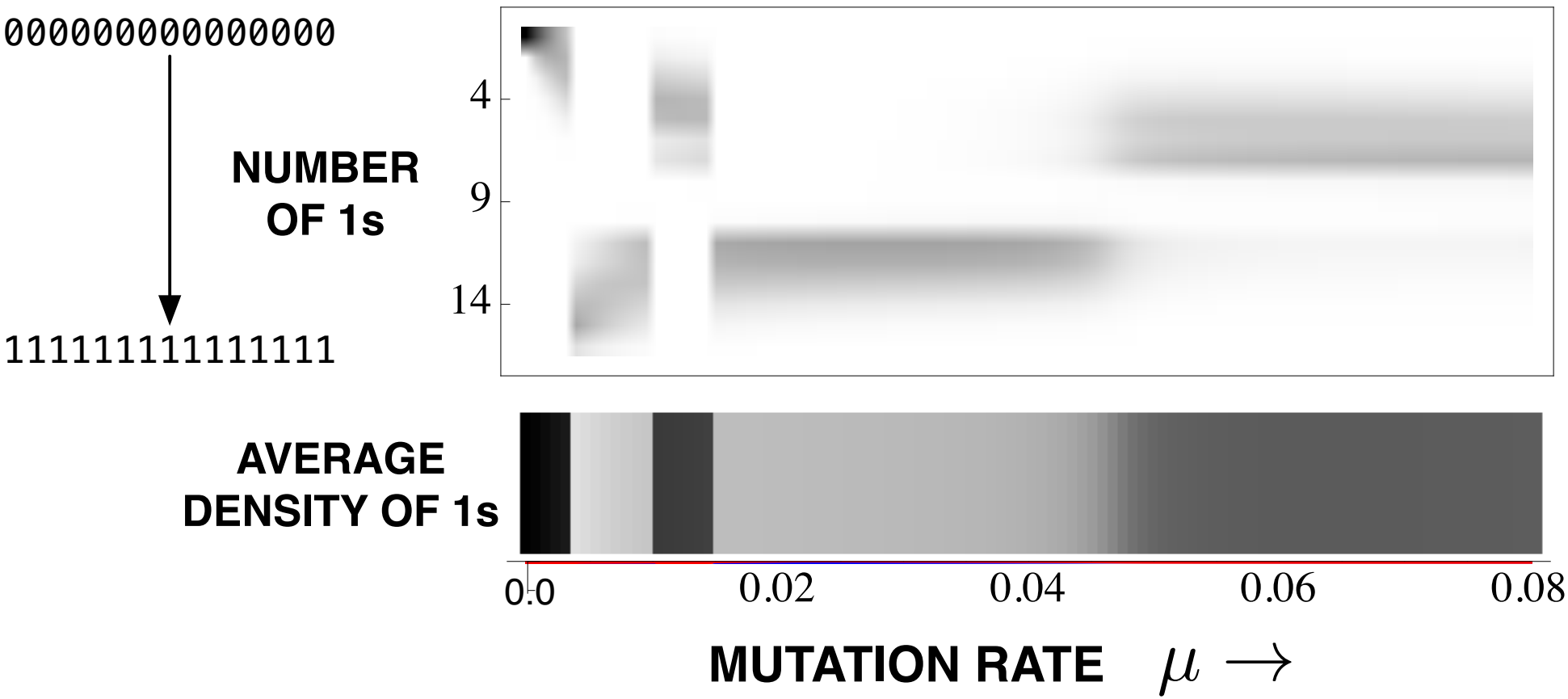
- Genotype fitness is a function of the number of mutations to the "master sequence" 000000000000000.
- 15 loci, 2 alleles per locus, i.i.d. mutation at rate μ per locus.



Result: Multiple, Reversing "Error Catastrophes" — A "Yo-Yo"

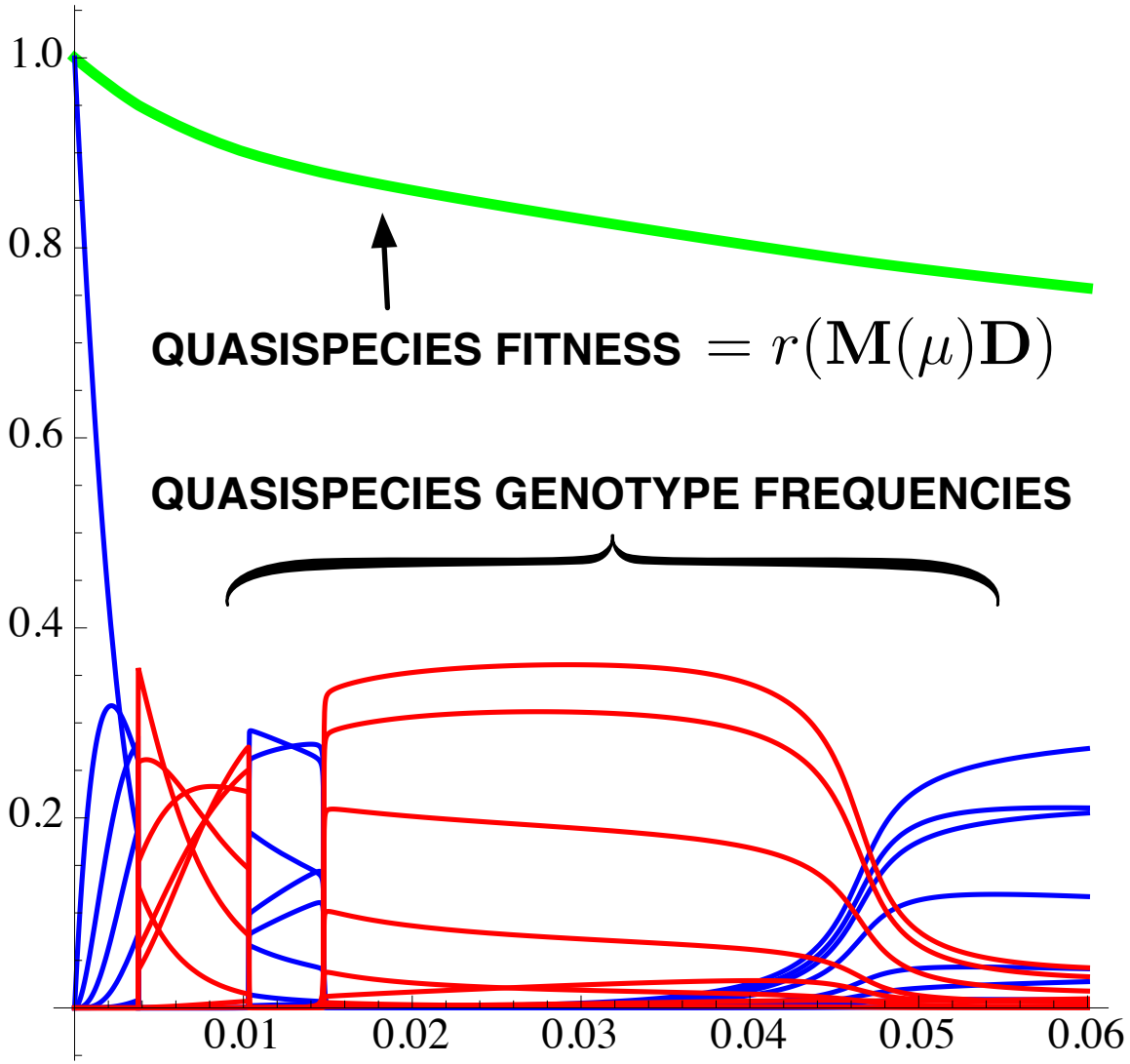


Result: Multiple, Reversing "Error Catastrophes" — A "Yo-Yo"

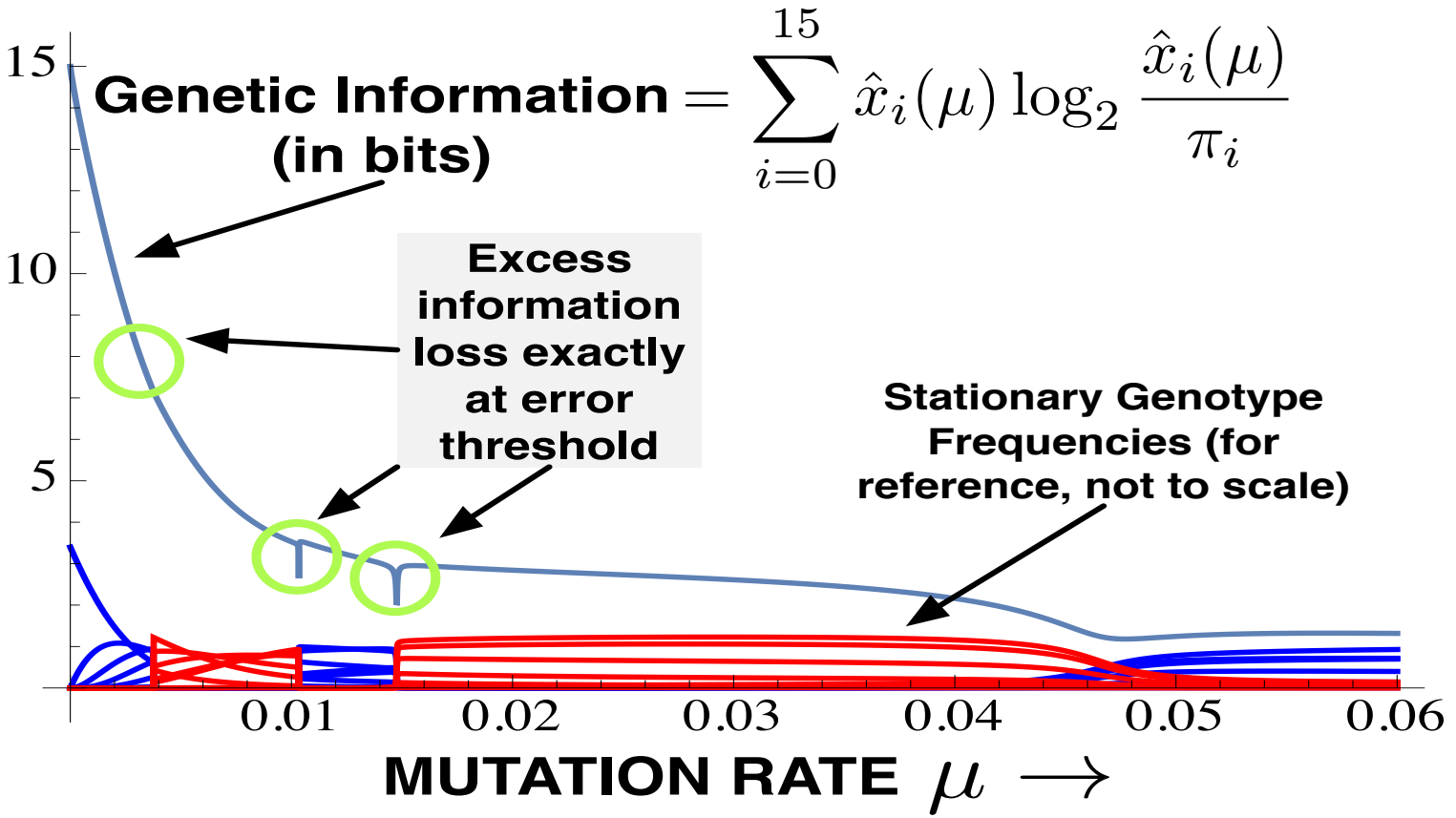


Fitness Declines Smoothly Despite Four Error Thresholds

- Drama in the genotype frequencies contrasts with
- Smooth decline of the quasispecies fitness $r(\mathbf{M}(\mu)\mathbf{D})$
- Illustrating theorems in A. (2011, 2015)



Genetic Information and Mutation Rates



- 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.

Conclusion 2

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.

Acknowledgements

Source: A. (2017) "Genetic Information, Mutation Rates, and the Lore of the Error Threshold"

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

A. 2011. An evolutionary reduction principle for mutation rates at multiple loci. *Bulletin of Mathematical Biology*, 73:1227–1270.

A. 2015. Fundamental properties of the evolution of mutational robustness. *arXiv preprint arXiv:1508.07866*.

A. 2017. Genetic information, mutation rates, and the lore of the error threshold. In *Proceedings of the 9th International Conference on Bioinformatics and Computational Biology (BICOB 2017)*, pages 223–230, Winona, MN, USA. The International Society for Computers and Their Applications (ISCA).

Arias, A., de Ávila, A. I., Sanz-Ramos, M., Agudo, R., Escarmís, C., and Domingo, E. 2013. Molecular dissection of a viral quasispecies under mutagenic treatment: positive correlation between fitness loss and mutational load. *Journal of General Virology*, 94(4):817–830.

Barbieri, M. 2012. Code biology–a new science of life. *Biosemiotics*, 5(3):411–437.

Eigen, M. and Schuster, P. 1977. The hypercycle: A principle of natural self-organization. *Naturwissenschaften*, 64:541–565.

Hermisson, J., Redner, O., Wagner, H., and Baake, E. 2002. Mutation–selection balance: Ancestry, load, and maximum principle. *Theoretical Population Biology*, 62(1):9–46.

References II

Krakauer, D. C. and Rockmore, D. N. 2015. The mathematics of adaptation (or the Ten Avatars of Vishnu). In Higham, N. J., editor, *The Princeton Companion to Applied Mathematics*. Princeton University Press, Princeton, NJ.

Li, D.-F., Cao, T.-G., Geng, J.-P., Gu, J.-Z., An, H.-L., and Zhan, Y. 2015. Statistical properties and error threshold of quasispecies on single-peak gaussian-distributed fitness landscapes. *Journal of theoretical biology*, 380:53–59.

Sarkisyan, K. S., Bolotin, D. A., Meer, M. V., Usmanova, D. R., Mishin, A. S., Sharonov, G. V., Ivankov, D. N., Bozhanova, N. G., Baranov, M. S., Soylemez, O., Bogatyreva, N. S., Vlasov, P. K., Egorov, E. S., Logacheva, M. D., Kondrashov, A. S., Chudakov, D. M., Putintseva, E. V., Mamedov, I. Z., Tawfik, D. S., Lukyanov, K. A., and Kondrashov, F. A. 2016. Local fitness landscape of the green fluorescent protein. *Nature*.

Schuster, P. 2013. The mathematics of Darwinian systems. In Eigen, M., editor, *From Strange Simplicity to Complex Familiarity: A Treatise on Matter, Information, Life, and Thought*, chapter Appendix A, pages 667–700. Oxford University Press.

Strelioff, C. C., Lenski, R. E., and Ofria, C. 2010. Evolutionary dynamics, epistatic interactions, and biological information. *Journal of Theoretical Biology*, 266(4):584–594.

Tripathi, K., Balagam, R., Vishnoi, N. K., and Dixit, N. M. 2012. Stochastic simulations suggest that HIV-1 survives close to its error threshold. *PLoS Computational Biology*, 8(9):e1002684.

References III

Woodcock, G. and Higgs, P. G. 1996. Population evolution on a multiplicative single-peak fitness landscape. *Journal of Theoretical Biology*, 179(1):61–73.

APPENDICES

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."

Krakauer and Rockmore (2015) The Mathematics of Adaptation (Or the Ten Avatars of Vishnu)

- "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 ."