

Okinawa Institute of Science and Technology

沖縄科学技術研究基盤整備機構

OIST

Lecture Two

Genome: Information Carrier at the “Edge of Chaos”

HC Lee

National Central University

Life at the edge of chaos

- Edge of chaos
 - Computational system
 - Cellular automata
 - Transition to criticality
- Life at the Edge of chaos
 - Life involves complex computation
 - Technical apparatus for description still missing
- Genome as Life
 - Chaos as a state of near randomness
 - Textual complexity of a genome represent computational ability
 - Dynamics of genome evolution

Terminology & Notations

- Consider genome with fractional **AT-content** p (then fractional GC-content $q=1-p$)
 - When $p>0.5$, there will be more AT-rich words than GC-rich words
- Partition k -letters words (k -mer) into sets, called m -sets S_m , $m=0,1,\dots,k$; elements in S_m are k -mers with m ATs, $\bigcup_m S_m = S$
- Total number of kinds of k -mers is $\tau=4^k$, of kinds of k -mers in m -set is τ_m . Let u be a k -mer,

$$L_m = \sum_{u \in S_m} f_u, \quad \sum_{u \in S} f_u = L - k + 1, \quad \sum_{u \in S} f_u = \sum_m L_m = L$$

$$\tau_m = \binom{k}{m} 2^k, \quad L_m^{\{\infty\}} = L \binom{k}{m} p^m q^{k-m},$$

$$\bar{f} = L/\tau, \quad \tau = 4^k, \quad \bar{f}_m^{\{\infty\}} = L_m^{\{\infty\}}/\tau_m$$

Shannon entropy (briefly)

- **Shannon entropy** for a system frequency set $\{f_i \mid \sum_i f_i = L\}$ or a spectrum $\{n_f\}$ is

$$H = - \sum_i f_i/L \log (f_i/L) = - \sum_f n_f f/L \log (f/L)$$

- Suppose there are τ types of events: $\sum_i = \tau$. Then H has **maximum value** when every f_i is equal to N/τ :

$$H_{max} = \log \tau$$

- For a genomic k -frequency set: $\tau = 4^k$, L = genome length.

$$H_{max} = 2k \log 2$$

Divergence in frequency distribution (D)

- Let D be coefficient of invariance (SD/mean) of distribution of frequency of occurrence of k -mers

$$D = (CV)^2 = \left(\frac{\sigma}{\bar{f}} \right)^2 = \frac{1}{\tau \bar{f}^2} \sum_{u \in \mathcal{G}_k} (f_u - \bar{f})^2$$

- Random sequence: $D \sim L^{-1/2}$
- Define l_{eff} for a genome to be the random sequence length having the genomic D
- Genomes have (almost) universal l_{eff} , about 150-600 bases long (for 2-mers)

Shannon information & relative spectral width

- **Shannon information:** information is *decrease* in H : define

$$R = \log \tau - H$$

Shannon called R/H_{max} redundancy; Gatlin (1972) called R divergence

- Relation to **relative spectral width** (for unimodal distribution)

$$\sigma = \Gamma/2 \bar{f}$$

$$R = \sigma^2/2 + O(\sigma^3)$$

- Shannon information and relative spectral width (“fluctuation part” from Lecture 1) are equivalent measures

HD Chen, et al. Divergence and Shannon information in genomes. [Phys. Rev. Lett. 94 \(2005\) 178103](#).

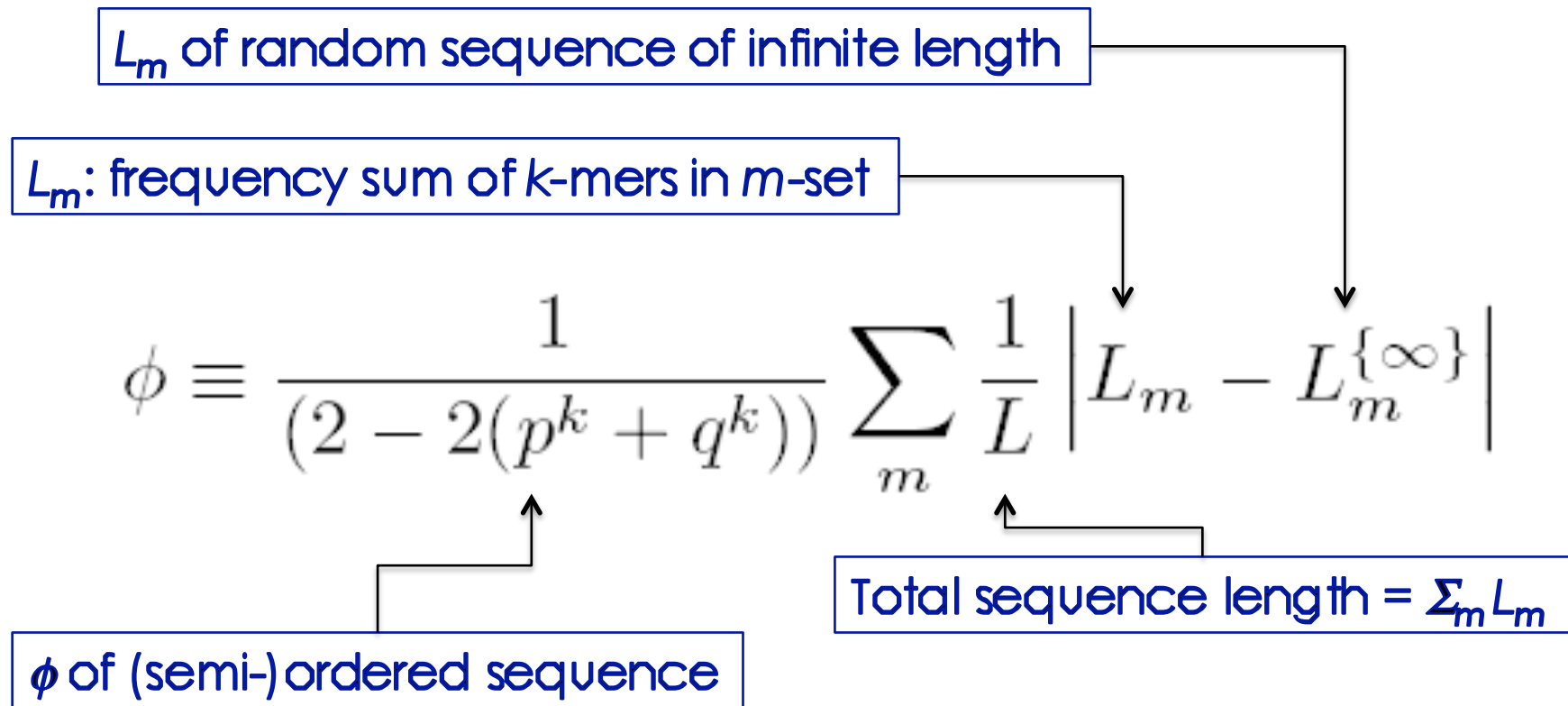
CH Chang, et al. Shannon Information in Complete Genomes. [J. Bioinfo. & Comp. Biology 3 \(2005\) 587-608](#).

$R = \log \tau - H$ is a good definition

Table 1: Shannon entropy H and information R in units of $\log 2$ in the k -spectra of the genome sequence of *P. aerophilum* and of the random sequence obtained by randomizing the genome. R_{ex} is the expected information in a random sequence. **Sequences have AT/CG= 50/50**

k	Random sequence			Genome sequence		$R_{\text{gen}}/R_{\text{ran}}$
	H	R	R_{ex}	H	R	
2	3.9999	5.90 E-6	5.77 E-6	3.973	2.66 E-2	4500
3	5.9999	3.72 E-5	3.46 E-5	5.933	6.65 E-2	1922
4	7.9999	1.72 E-4	1.62 E-4	7.881	1.18 E-1	728
5	9.9993	7.26 E-4	7.53 E-4	9.821	1.79 E-1	246
6	11.999	2.94 E-3	2.90 E-3	11.75	2.74 E-1	94
7	13.988	1.18 E-3	1.17 E-3	13.66	3.35 E-1	29
8	15.955	4.78 E-2	4.71 E-2	15.53	4.69 E-1	10
9	17.798	2.02 E-1	1.88 E-1	17.26	7.33 E-1	3.0
10	19.xxx	x.xx E-1	5.24 E-1	19.xx	x.xx E-1	-

An Order Index ϕ



Note: ϕ is a measure of differential in averages

ϕ is a measure of order in a sequence

- An (semi-)ordered sequence

AT...TATTATATTAATATTTAGCCGGGGCGGGCGC...GG

or a checker-board sequence

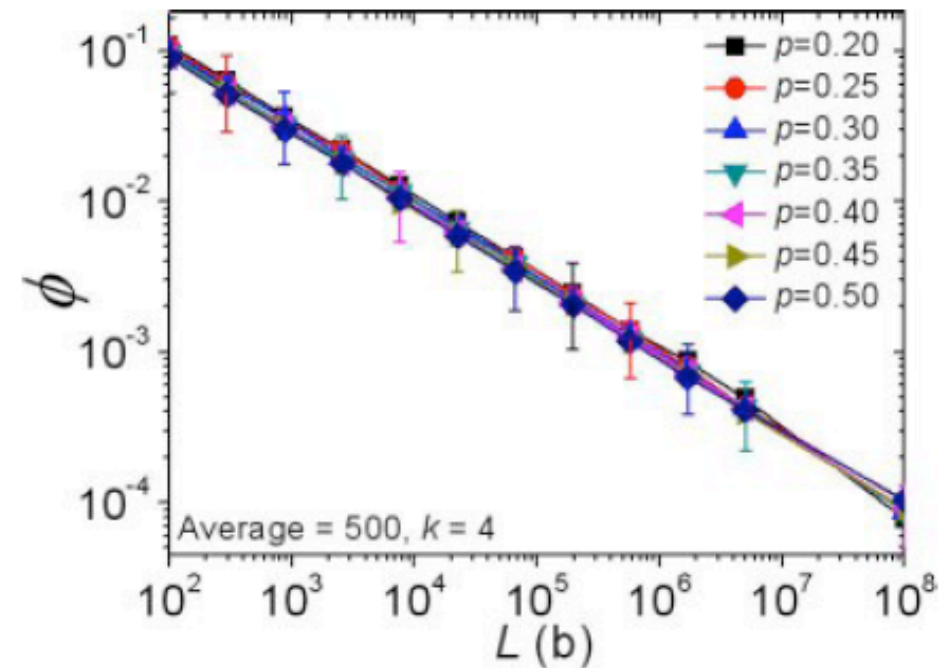
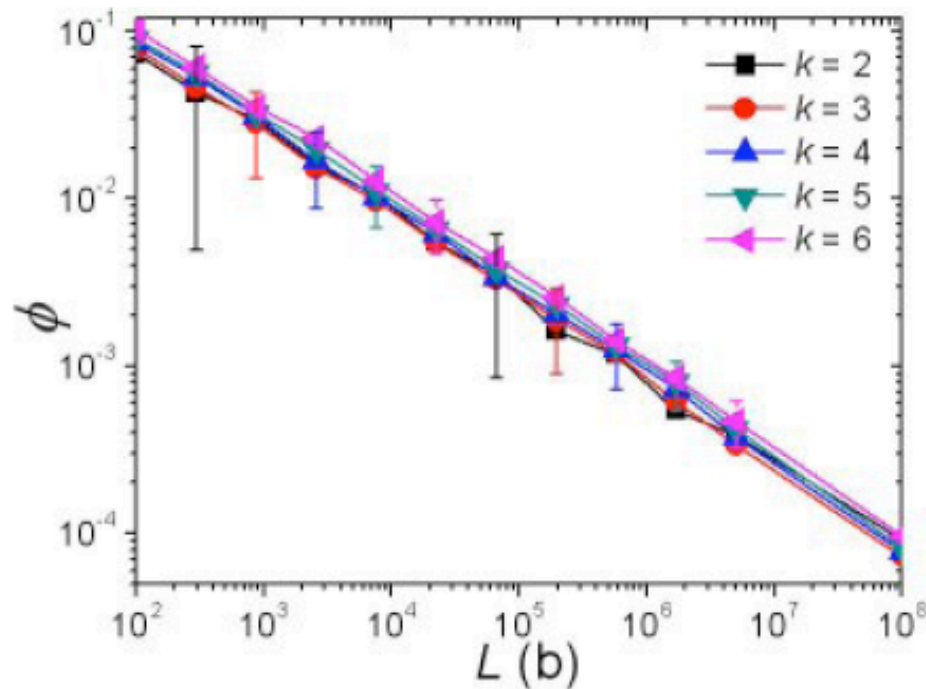
...AGAGTGACAGTCTGTCTCACTG...

have $\phi \sim 1$

- A random sequence has

$$\phi \sim L^{-1/2} \sim 0 \text{ for large } L$$

ϕ scales as $L^{-1/2}$ for random sequences



Depends only weakly on k or AT-content (p).

Averaged over k and p :

$$\phi^{\{ran\}}(k; p) = c_{\phi} L^{-\gamma_{\phi}}, \quad \gamma_{\phi} = 0.500 \pm 0.005, \quad c_{\phi} = 1.0 \pm 0.2$$

An equivalent length L_ϕ for order index

Use the relation

$$\phi\{ran\} \approx L^{-1/2}$$

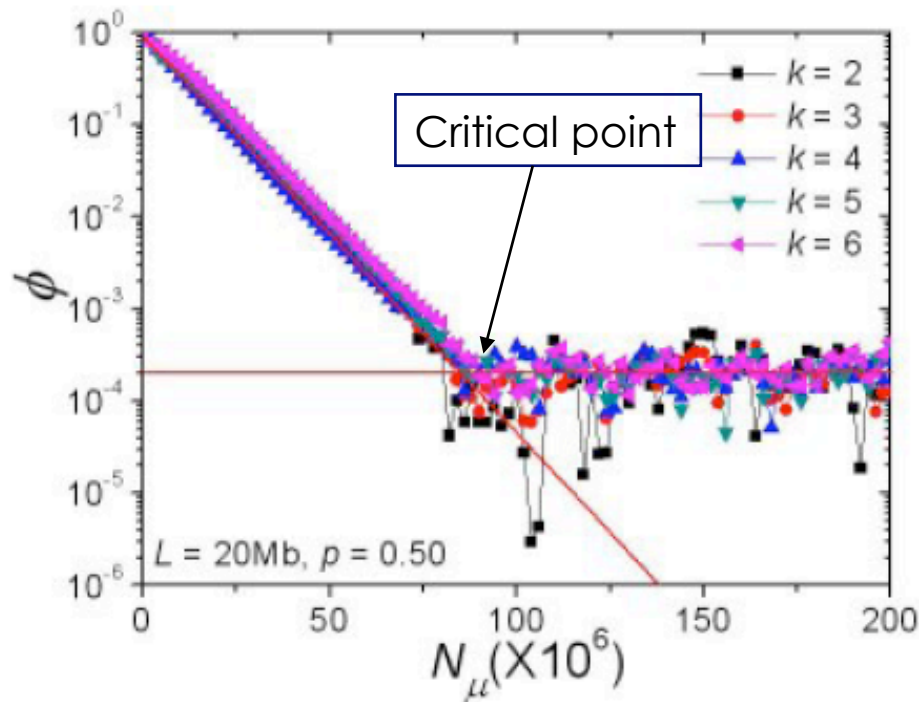
to define an (order-index) **Equivalent Length** for a ϕ -valued sequence:

$$L_\phi(\phi) = \phi^{-2} ,$$

the nominal length of a (non-random) sequence whose order index is ϕ .

(Note. Unlike the CV-defined and k -dependent L_e , L_ϕ is essentially independent of k .)

ϕ decreases exponentially with increasing number of point mutations



For an ordered sequence, ϕ drops exponentially from $\phi=1$ as a function of N_μ until it reaches a critical point when the sequence has become random.

N_μ : # of random mutations

$$\phi = \begin{cases} \exp(-2N_\mu/L), & N_\mu \lesssim N_{\mu c}; \\ \phi_c \approx L^{-1/2}, & N_\mu > N_{\mu c} \end{cases}$$

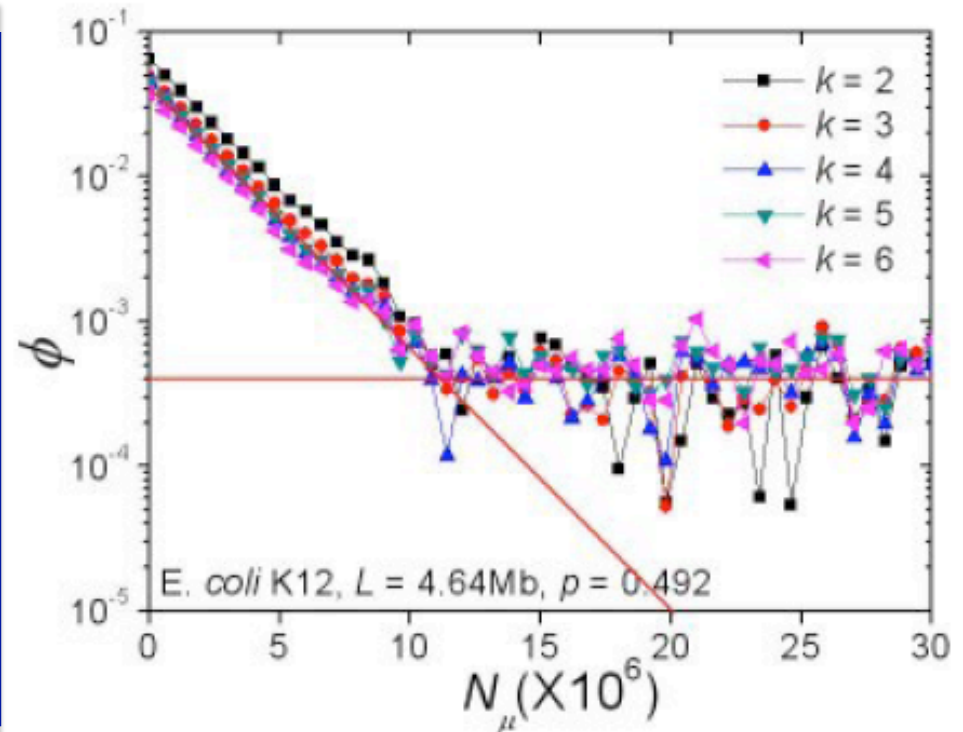
ϕ decreases exponentially with increasing number of point mutations

If sequence already has $\phi_0 < 1$, then as a function of $N\mu$

$$\phi = \phi_0 \exp(-2N_\mu/L),$$

until ϕ reaches its critical value $\phi_c \sim L^{-1/2}$.

Let $\mu = N_\mu/L$ be mutation density (mutation per site)



- **Equivalent mutation density** for a ϕ -valued sequence:

$\mu_{eq}(\phi) \equiv \ln \phi^{-1/2}$, the nominal mutation density needed to bring an ordered sequence to a state of ϕ

The critical mutation density

Given

$$\phi = \begin{cases} \exp(-2N_\mu/L), & N_\mu \lesssim N_{\mu c}; \\ \phi_c \approx L^{-1/2}, & N_\mu > N_{\mu c} \end{cases}$$

The critical mutation density that will bring an ordered sequence to a state of randomness is given by

$$\phi_c = \exp(-2N_\mu/L) = \exp(-2\mu_c) = L^{-1/2}$$

That is:

$$\mu_c = (1/4) \ln L$$

(For $L = 10^6$ to 10^8 ,
 $\mu_c = 3.4$ to $4.6/\text{site}$)

Equivalent mutation density is a path-independent state quantity

- Want to test whether ϕ (or equivalently, μ_{eq}) is akin to a potential energy, or a quantity that defines a state, **but not how that state is arrived at**.
 - The 4.6 Mb *E. coli* DID NOT arrive at its present state by random point mutation from an ordered sequence.
 - It is measured to have $\phi = 0.049$, or $\mu_{eq} = 1.5/\text{site}$.
 - The critical μ_{eq} for a 4.6 Mb sequence is $\mu_c = (1/4) \ln(4.6 \times 10^6) = 3.8/\text{site}$.
 - If we assume ϕ is a path-independent state quantity, then we **predict** it will take $(\mu_c - \mu_{eq}) \times L = 1.1 \times 10^7$ additional mutations to randomize *E. coli*
 - The actual number is found to be $1.1(\pm 0.1) \times 10^7$

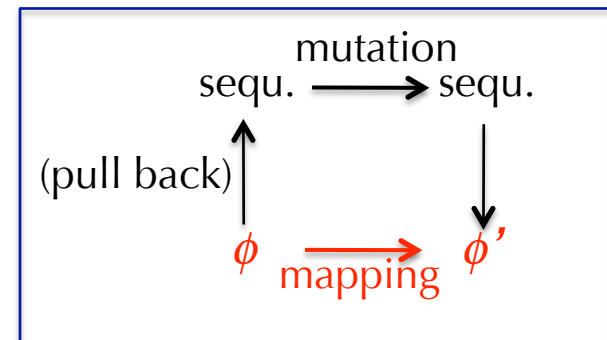
State of randomness as a fixed point

Considered as a dynamical system driven by mutations, the state of randomness is a **fixed point** in ϕ space.

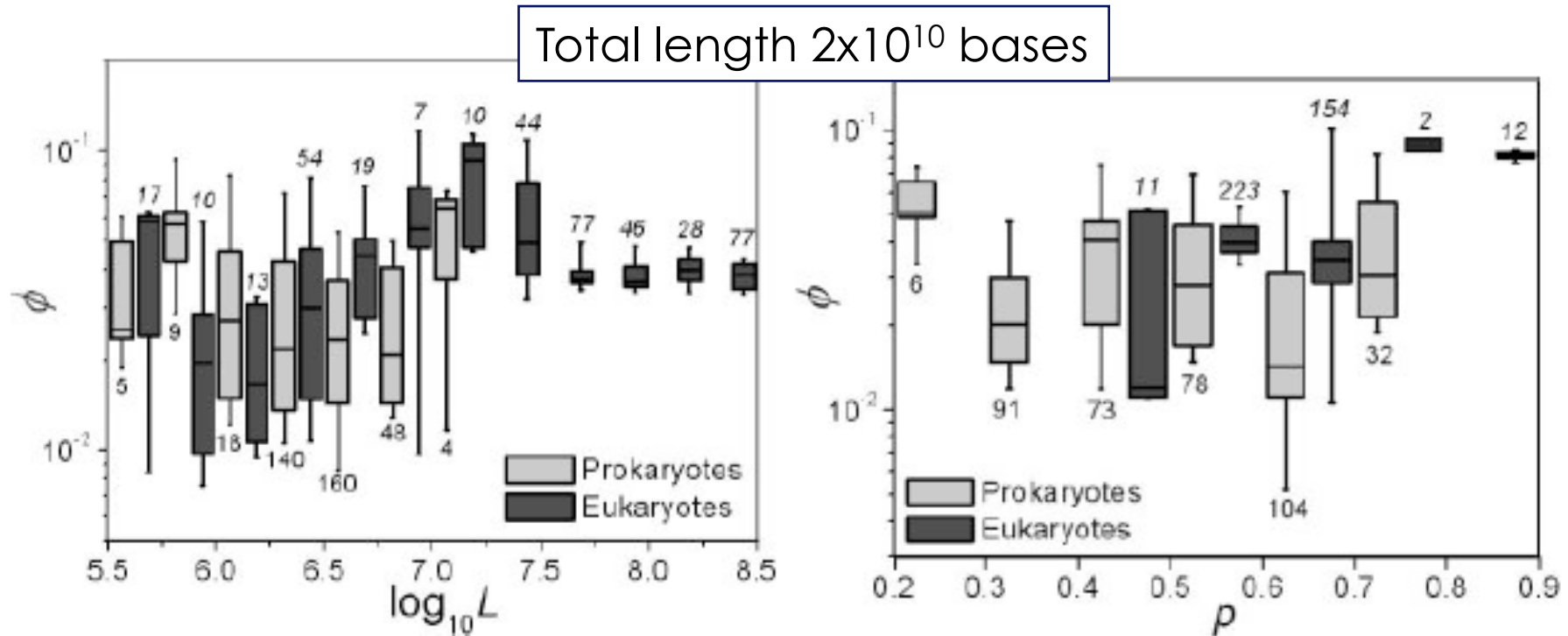
[Fixed point. Consider a function $f(x)$ mapping a point x in a space to another point x' in the same space. Then x_* is a fixed point of f if

$$f(x_*) = x_*.]$$

Here the action causing the mapping is mutation, and the space is the ϕ -space. Mutation takes the sequence from one ϕ to another ϕ . When the sequence is random, mutation maps ϕ_c back to itself.



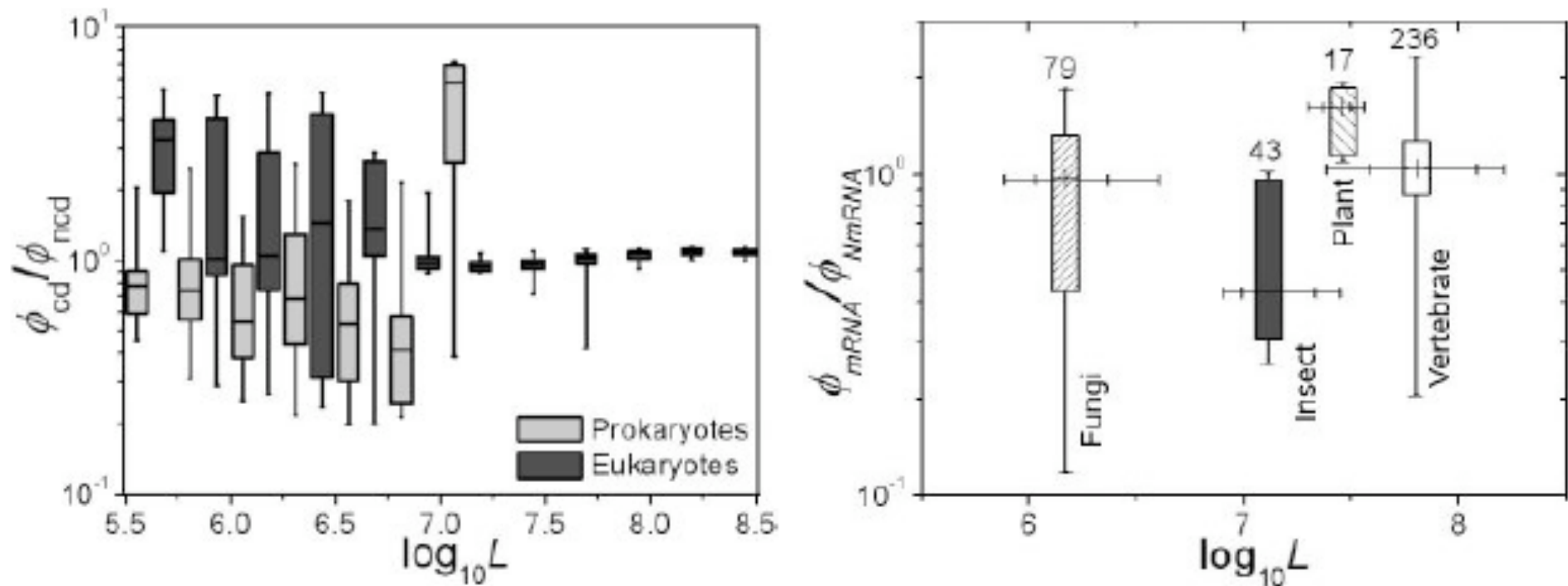
For ~800 complete genomes extant in GenBank, ϕ is essentially length- and base-composition-independent



- Genomic ϕ congregates in a narrow range

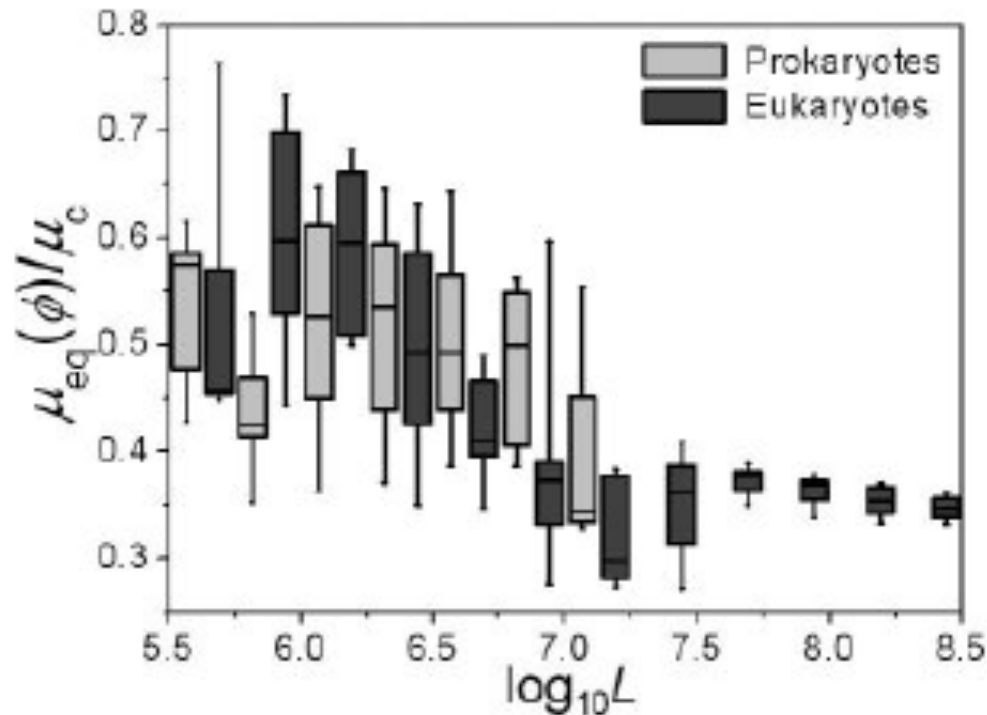
$$\phi_g \equiv 0.037 \pm 0.027$$

Coding (genic) and non-coding parts have similar ϕ



- Dynamics of genome evolution leading ϕ to ϕ_g is not under strong (genic) selection pressure
- Predominant characteristics is neutral

Genomes are half as random as random sequences



$$\mu_{eq}(\phi)/\mu_c \sim 0.45 \pm 0.11$$

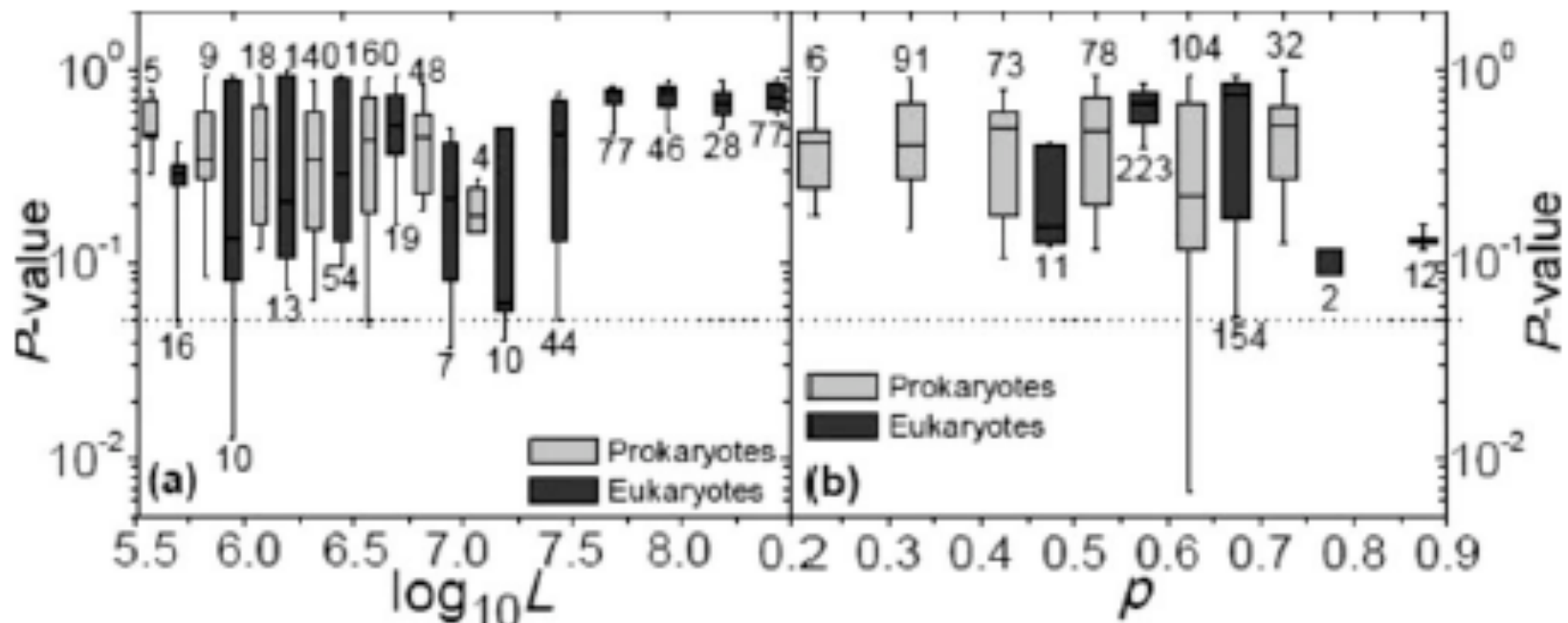
- $\mu_{eq} \sim 1.8 \text{ b}^{-1}$ implies a genome is as random as an ordered sequence becomes after each site has on average been mutated 1.8 times.
- Genome is **at the Edge of Chaos**

Genomes form a universality class in ϕ

Vast majority of genomes have

$$\ln \phi_g = -3.49 \pm 0.65, \text{ or } \phi_g = 0.031^{+0.028}_{-0.015}$$

P-value of genomes **belonging** to the universality class define by ϕ_g



37 (out of ~800) chromosomes belong to universality class in ϕ with $P < 0.05$

TABLE II. Chromosomes belonging to the universal set defined by Eq. (8) with $P < 0.05$.

Name	Accession no. ^a	$\bar{\phi}$ ^b	P value ^b	Name	Accession no. ^a	$\bar{\phi}$ ^b	P value ^b
<i>S. aureus</i>	9 strains ^c	$\sim 4.4(-3)$	$\sim 3.0(-3)$	<i>A. marginale</i>	4842	$4.45(-3)$	$3.15(-3)$
<i>S. epidermidis</i>	4461	$4.87(-3)$	$4.89(-3)$	<i>C. felis</i>	7899	$5.20(-3)$	$6.66(-3)$
<i>L. johnsonii</i>	5362	$5.58(-3)$	$9.18(-3)$	<i>S. hemolyticus</i>	7168	$5.79(-3)$	$1.08(-2)$
<i>S. epidermidis</i>	2976	$6.49(-3)$	$1.77(-2)$	<i>M. mobile 163 K</i>	6908	$6.80(-3)$	$2.14(-2)$
<i>T. denitrificans</i>	7404	$7.12(-3)$	$2.58(-2)$	<i>L. acidophilus</i>	6814	$7.34(-3)$	$2.90(-2)$
<i>G. sulfurreducens</i>	2939	$7.40(-3)$	$2.99(-2)$	<i>F. tularensis</i>	7880	$7.50(-3)$	$3.15(-2)$
<i>W. succinogenes</i>	5090	$7.51(-3)$	$3.17(-2)$	<i>C. hydrogenoformans</i>	7503	$1.23(-1)$	$3.20(-2)$
<i>M. hungatei</i>	7796	$7.75(-3)$	$3.57(-2)$	<i>F. tularensis</i>	6570	$7.90(-3)$	$3.84(-2)$
<i>C. caviae</i>	3361	$7.94(-3)$	$3.91(-2)$	<i>M. succiniciproducens</i>	6300	$1.15(-1)$	$4.04(-2)$
<i>C. abortus</i>	4552	$8.06(-3)$	$4.14(-2)$	<i>X. fastidiosa 9a5c</i>	2488	$8.12(-3)$	$4.25(-2)$
<i>P. marinus</i>	7335	$8.19(-3)$	$4.37(-2)$	<i>S. tokodaii</i>	3106	$8.47(-3)$	$4.96(-2)$
<i>S. cerevisiae</i>	Chr V	$6.00(-3)$	$1.26(-2)$	<i>S. cerevisiae</i>	Chrs XV, III	$\sim 7.7(-3)$	$\sim 3.5(-2)$
<i>S. cerevisiae</i>	Chr VI	$8.43(-3)$	$4.87(-2)$	<i>A. mellifera</i>	8 chrs. ^d	$\sim 1.1(-1)$	$\sim 4.8(-2)$

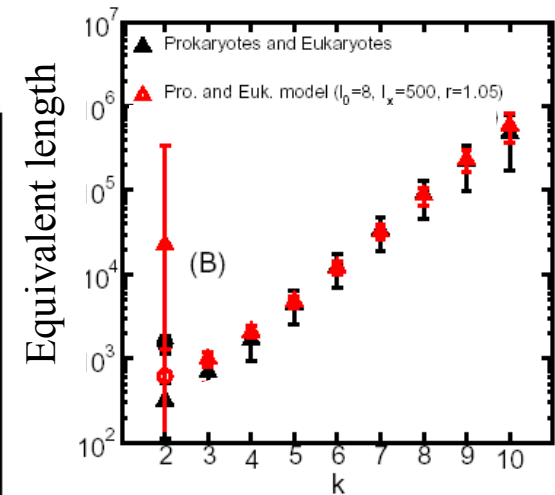
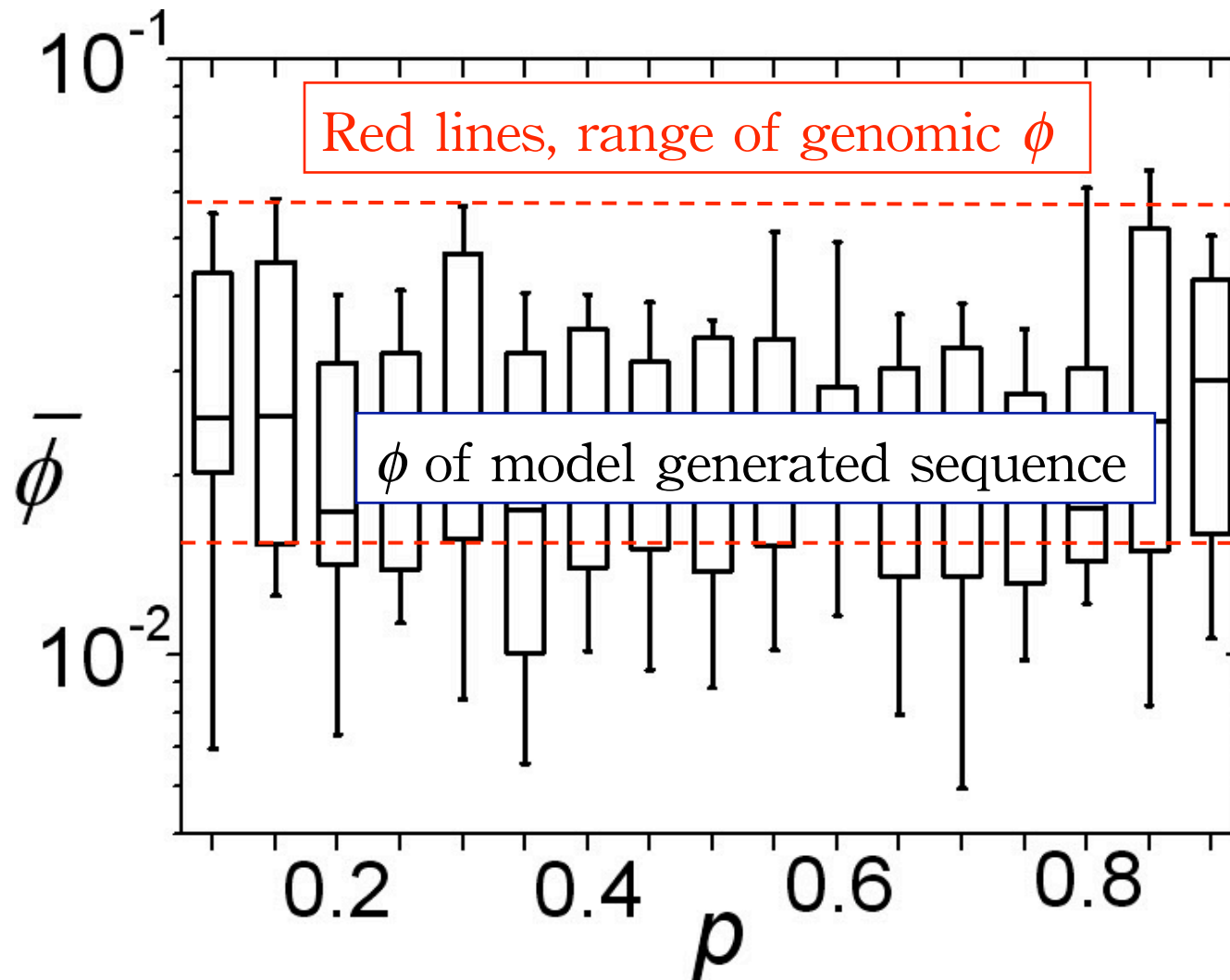
^a4842 indicates the accession no. NC_004842.

^bThe value $4.4(-3)$ means 4.4×10^{-3} .

^cThe nine strains, in order of increasing P value, are 3923, 2953, 7793, 7795, 2952, 7622, 2951, 2758, and 2745.

^dThe eight chromosomes, in order of increasing P value, are XV, X, XII, II, IV, V, I, and XI.

Artificial sequence with genomic L_e
generated in RSD model has genomic ϕ



An empirical function of information capacitance

Define an “information capacity” function

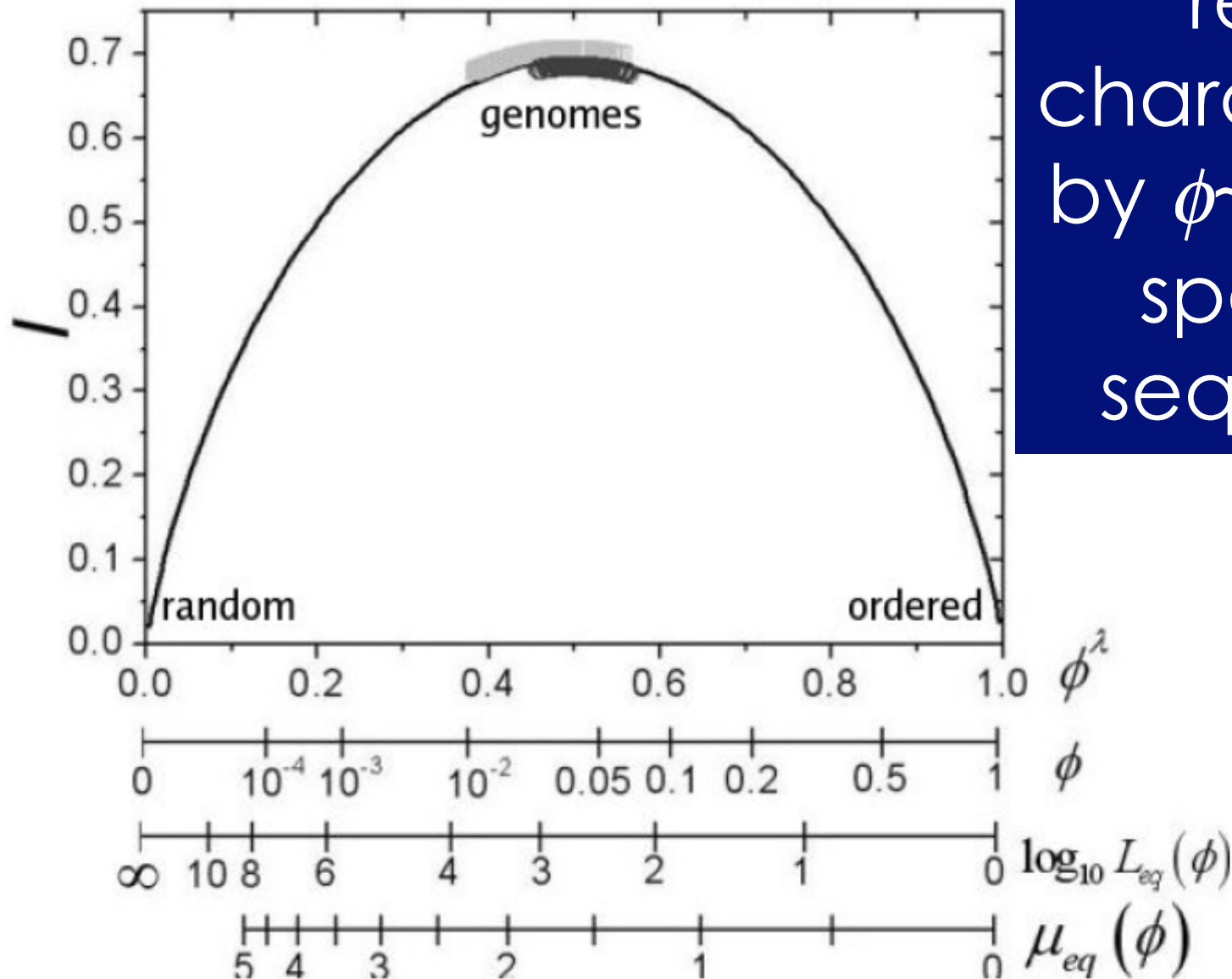
$I(z)$ such that: the variable z is a scaling function of ϕ with $z|_{\phi=0}=0$ and $z|_{\phi=1}=1$, and I has two minima at $I(0)=I(1)=0$ and a maximum at $z|_{\phi=\phi_g}=0.5$. The simplest solution is

$$I(z) = -z \ln z - (1 - z) \ln(1 - z); \quad z = \phi^{0.21}. \quad (10)$$

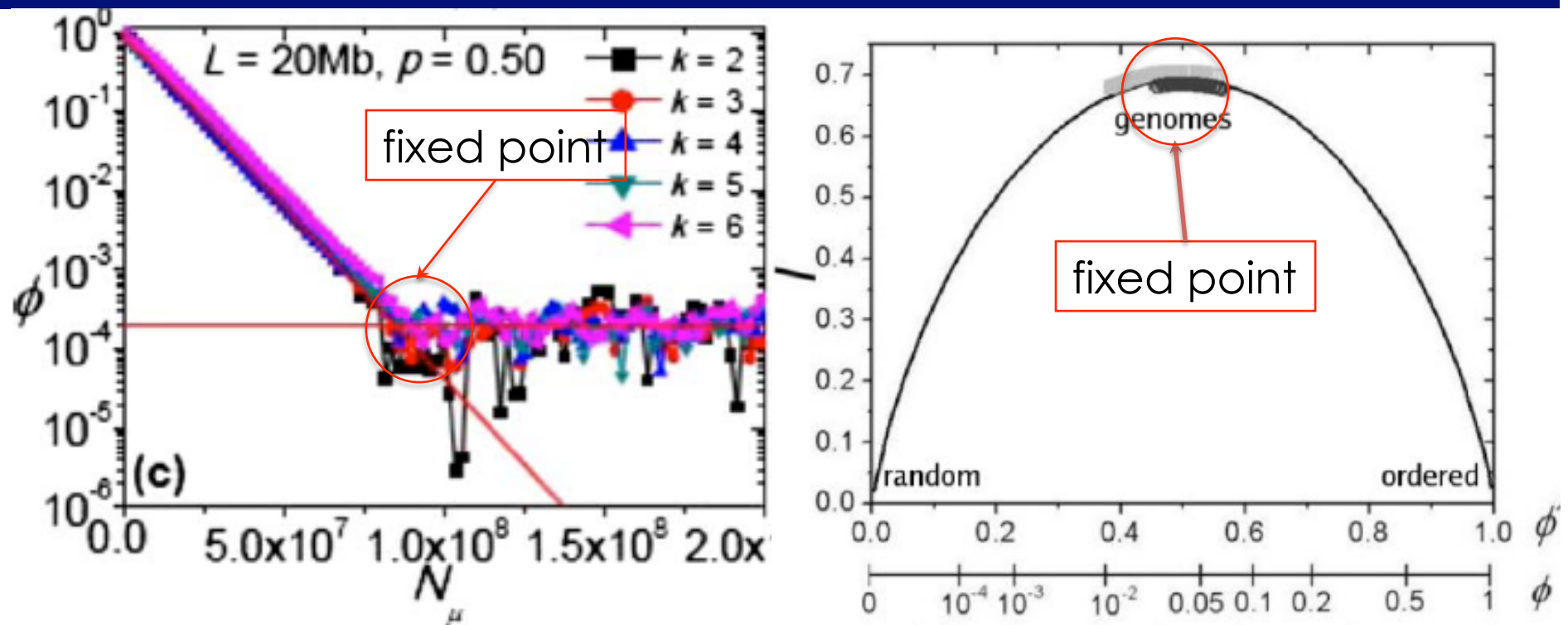
Say “information capacitance”, not “information”, because ϕ_g is universal but sequence length is not; not “information density”, because not every sequence with ϕ_g has information.

Genomes reside in a small distinct

region
characterized
by $\phi \sim \phi_g$ in the
space of
sequences



Genomes at a fixed point

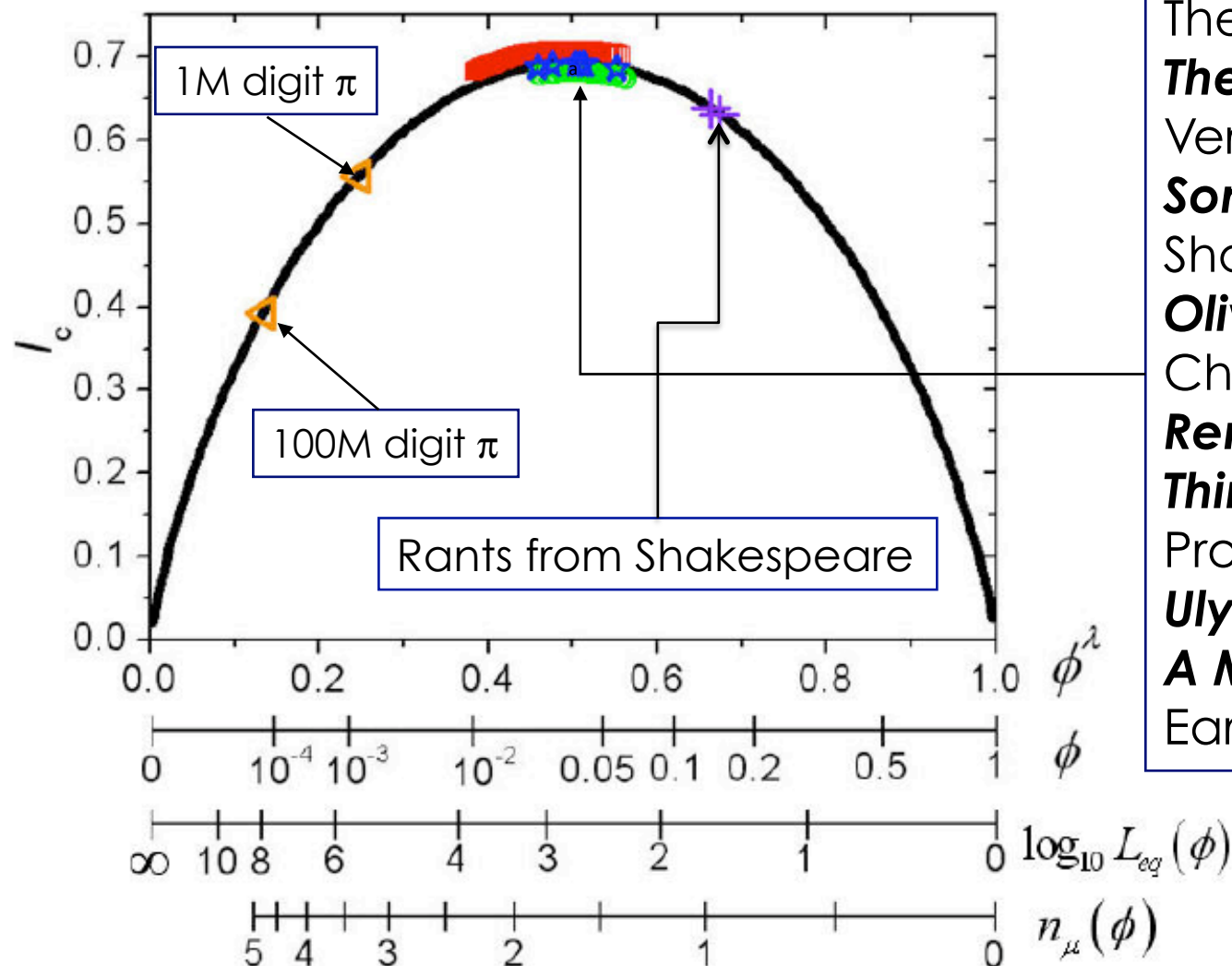


Recall: Sequences are driven by random mutations to a state-of-randomness fixed point. Similarly, genomes are driven by the dynamics of a robust evolutionary process (yet to be identified) to a **fixed-point** ϕ_g .

The two types of fixed points are driven by different dynamics

- Random sequence fixed point
 - $\phi_c \sim L^{-1/2}$, depends on sequence length
 - Driven by random point mutation
- Genome fixed point
 - $\phi_c = 0.016-0.059$ is universal (and independent on length)
 - Driven by “robust evolution process”
 - Our guess: random (+plus tandem) segmental duplication + plus point mutation

The $\phi \sim \phi_g$ “fixed point” shared by literature classics

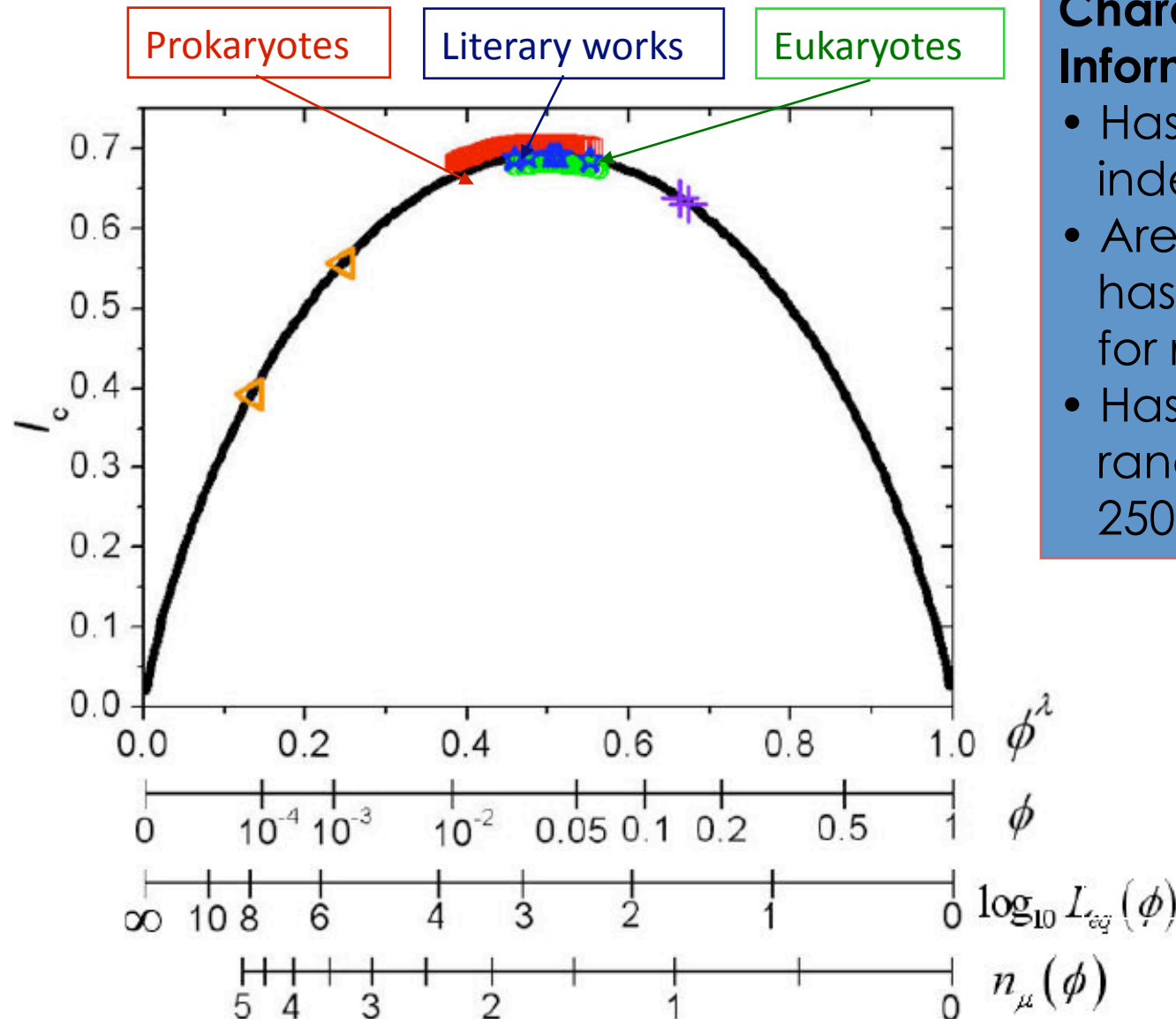


The six classics:
The Bible, King James Version;
Sonnets, William Shakespeare;
Oliver Twist, Charles Dickens;
Remembrance of Things Past, Marcel Proust;
Ulysses, James Joyce;
A Moveable Feast, Earnest Hemingway

Classics, rants, and π

- Making pseudogenomes of classics
 - (adjlsy) to A; (chiopq) to C; (efgnvxc) to G; (bkmrtuw) to T
 - All six classics have $p_A \sim p_C \sim p_G \sim p_T = 0.250 \pm 0.007$, or $p \sim 0.50 \pm 0.02$.
- The rants (repeated 1M times)
 - “Though this be madness, yet there is method in’t” (Hamlet)
 - “All the perfumes in Arabia will not sweeten this hand” (MacBeth)
- π : equivalent length close to true length
 - Highly complex sequence, yet low information content.

Information capacitance



Characteristics of Information carriers:

- Has “maximum” I_c independent of length
- Are quasi-random – has $\frac{1}{2}$ mut'ns needed for randomization
- Has ϕ equivalent to random sequence 250 – 10000 b

Order index

Equivalent length

Equivalent Mutation density

Conjectural Inferences

- $\phi \sim \phi_g$ are high information capacitance states
- The observed shortness of L_ϕ suggests that the neutral process is dominated by (fixed, hence non-deleterious) segmental duplications
- No difference in coding and non-coding part suggest process is random/neutral
 - Random: low free-energy, easy access
- Random process can only built infrastructure, not information; actual information is acquired in mostly fitness driven point mutation events
 - Selective: difficult to access

A two-step genome growth

- Genome growth by a two-step process:
 - One neutral, robust, infrastructure-building and universal
 - The other selective, fine-tuning, information-gathering and diverse
 - Example: paradigm of accidental gene duplication followed by mutation driven subfunctionalization
- The twin-processes acted in a ratchet-like, complementary manner, driving the genome, in successive stages, to a state of maximum information capacity, and helping it to acquire, at each stage, near-maximum information content.

End of Lecture Two