Looking at Whole Genomes – Information, growth & evolution

System Biology Workshop
19-21 June, 2004, Taipei

HC Lee
Computational Biology Lab
Dept. Physics & Dept. Life Sciences
National Central University &
National Center for Theoretical Sciences

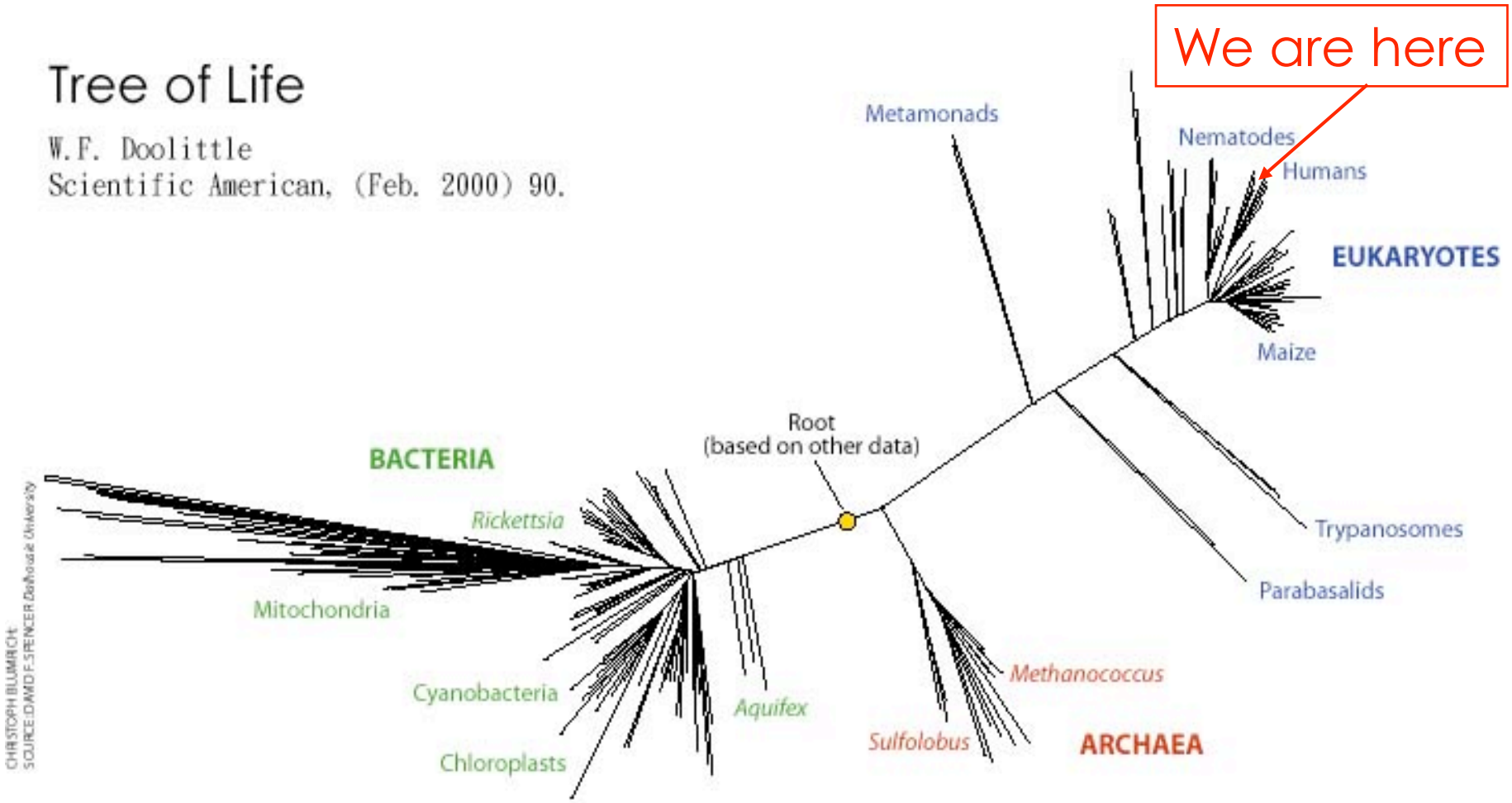
Plan of talk

- Genome and Life
- How did genome grow (life evolve) so quickly
- Textual spectral width & Shannon information
- Universality class of genomes
- Model for genome growth
- Self-similarity & randomness
- Substitution and duplication rates
- Discussion – implication in biology & evolution

Genome and Life

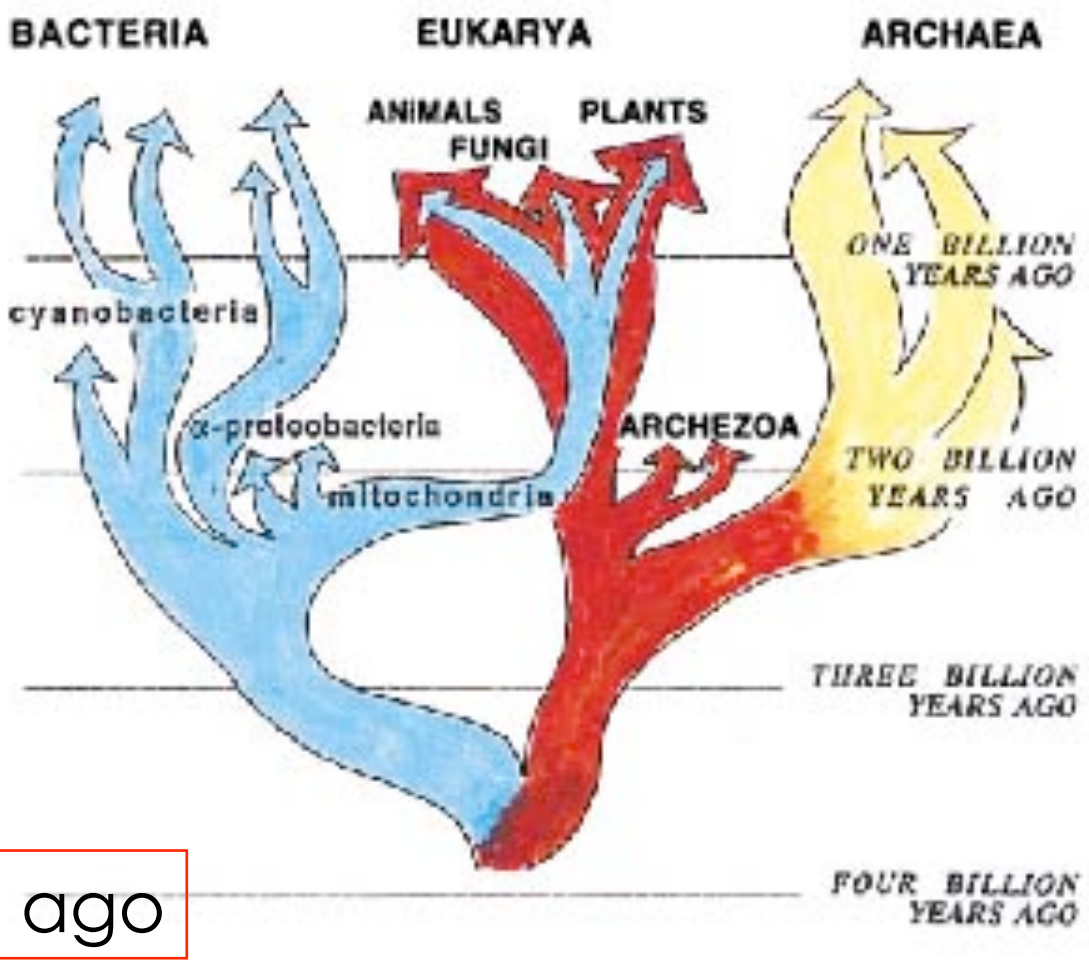
~ Life is the splendid expression genome – the ultimate organization of information

Life is highly diverse and complex



And it took a long time to get here

Divergence of species
W.F. Doolittle, PNAS 94 (1997) 12751.



now

4 billion yrs ago

Two approaches to Life Science

1. Local - "Biology"
   - Individual, specificity, uniqueness
2. Global - "Physics/Math/Stats"
   - Class, generality, universality, model

Today we take the GLOBAL route

Genome Growth, Entropy & Second law of Thermodynamics

~ How did genomes generate information stochastically

Evolution of Genomes and the Second Law of Thermodynamics

- **Genomes**
  - Grew and evolved (mainly) **stochastically**, modulated by natural selection
  - Bigger genomes carry more information than smaller ones
- **The second law of thermodynamics**:
  - the entropy of closed system can never decrease
  - a system that grows stochastically tends to acquire entropy
  - Increased randomness → more entropy
- **Shannon information**
  - Information decreases with increasing entropy

How did evolution fight against the Second Law?

- Genomes are not closed systems, but the 2nd law does make it difficult for the genome to simultaneously:
  - **grow stochastically**
  - **acquire more information**
    - lose entropy
    - gain order
- We propose an answer to this question

Genomes as Text - Spectral Width & Shannon Information

~ genomes have far more information than random sequences

Genomes are BIG

A stretch of genome from the X chromo-some of Homo sapien

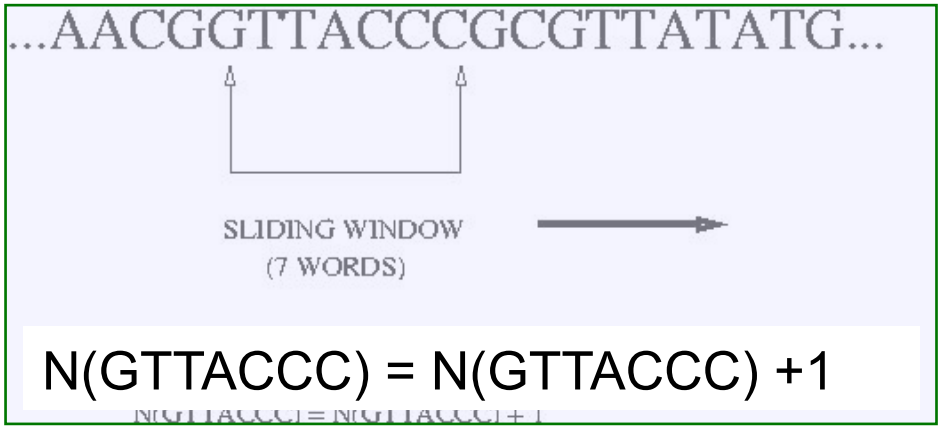
http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?val=2276452&db=Nucleotide&dopt=GenBank

The complete genome has 2,000,000 such Pages

```
   1 tgctgagaaa acatcaagctg tgtttctcct tccccaaag acacttcgca gcccctcttg
  61 ggatccagcg cagcgcaagg taagccagat gcctctgctg ttgccctccc tgtgggcctg
 121 ctctcctcac gccggccccc acctgggcca cctgtggcac ctgccaggag gctgagctgc
 181 aaaccccaat gaggggcagg tgctcccgga gacctgcttc ccacacgccc atcgttctgc
 241 ccccggcttt gagttctccc aggccccctct gtgcacccct ccctagcagg aacatgccgt
 301 ctgccccctt gagctttgca aggtctcggt gataatagga aggtctttgc cttgcaggga
 361 gaatgagtca tccgtgctcc ctccgagggg gattctggag tccacagtaa ttgcagggct
 421 gacactctgc cctgcaccgg gcgccccagc tcctccccac ctccctcctc catccctgtc
 481 tccggctatt aagacggggc gctcaggggc ctgtaactgg ggaaggtata cccgccctgc
 541 agaggtggac cctgtctgtt ttgatttctg ttccatgtcc aaggcaggac atgaccctgt
 601 tttggaatgc tgatttatgg attttccagg ccactgtgcc ccagatacaa ttttctctga
 661 cattaagaat acgtagagaa ctaaatgcat tttcttctta aaaaaaaaaa aaaccaaaaa
 721 aaaaaaaaaa aaaccaaaaa actgtactta ataagatcca tgcctataag acaaaggaac
 781 acctcttgtc atatatgtgg gacctcgggc agcgtgtgaa agtttacttg cagtttgcag
 841 taaaatgaca aagctaacac ctggcgtgga caatcttacc tagctatgct ctccaaaatg
 901 tatttttct aatctgggca acaatggtgc catctcggtt cactgcaacc tccgcttccc
 961 aggttcaagc gattctccgg cctcagcctc ccaagtagct gggaggacag gcacccgcca
1021 tgatgcccgg ttaatttttg tatttttagc agagatgggt tttcgccatg ttggccaggc
1081 tggtctcgaa ctcctgacct caggtgatcc gcctgccttg gcctcccaaa gtgctgggat
1141 gacaggcgtg agccaccgcg cccagccagg aatctatgca tttgcctttg aatattagcc
1201 tccactgccc catcagcaaa aggcaaaaca ggttaccagc ctcccgccac ccctgaagaa
1261 taattgtgaa aaaatgtgga attagcaaca tgttggcagg atttttgctg aggttataag
1321 ccacttcctt catctgggtc tgagcttttt tgtattcggt cttaccattc gttggttctg
1381 tagttcatgt ttcaaaaatg cagcctcaga gactgcaagc cgctgagtca aatacaaata
1441 gatttttaaa gtgtatttat tttaaacaaa aaataaaatc acacataaga taaaacaaaa
1501 cgaaactgac tttatacagt aaaataaacg atgcctgggc acagtggctc acgcctgtca
```

Genome as text - Frequencies of k-mers

- Genome is a text of four letters – A,C,G,T
- Frequencies of k-mers characterize the whole genome
  - E.g. counting frequencies of 7-mers with a "sliding window"
  - Frequency set $\{f_i \mid i=1 \text{ to } 4^k\}$



“Portrait” of genome and chaos game

A “Chaos Game”

For k-mers, 2^k by 2^k pixels, one spot gives color-code frequency of occurrence for each k-mer

Has “fractals”

BL Hao, HCL & SY Zhang
Chaos, Solitons & Fractals, 11 (2000) 825-836.

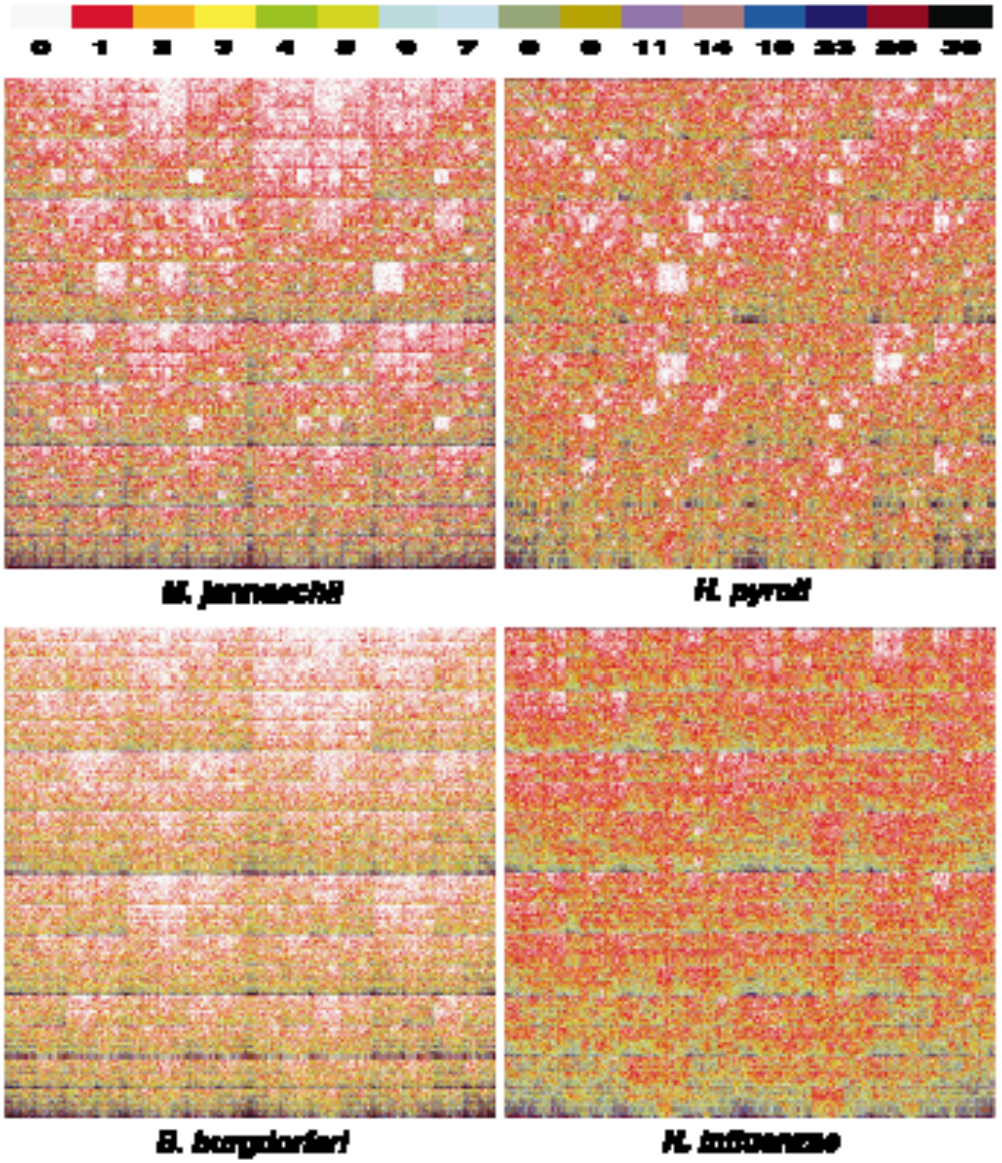


Fig. 3. Portraits of another 4 bacteria in $K = 9$ frames.

Prominent pattern of portrait determined by frequency of short oligonucleotides (words). (1) low CTAG; (2) A+T-rich; (3) AT-rich & high AC, CA, GT, TG; (4) high AA, TT.

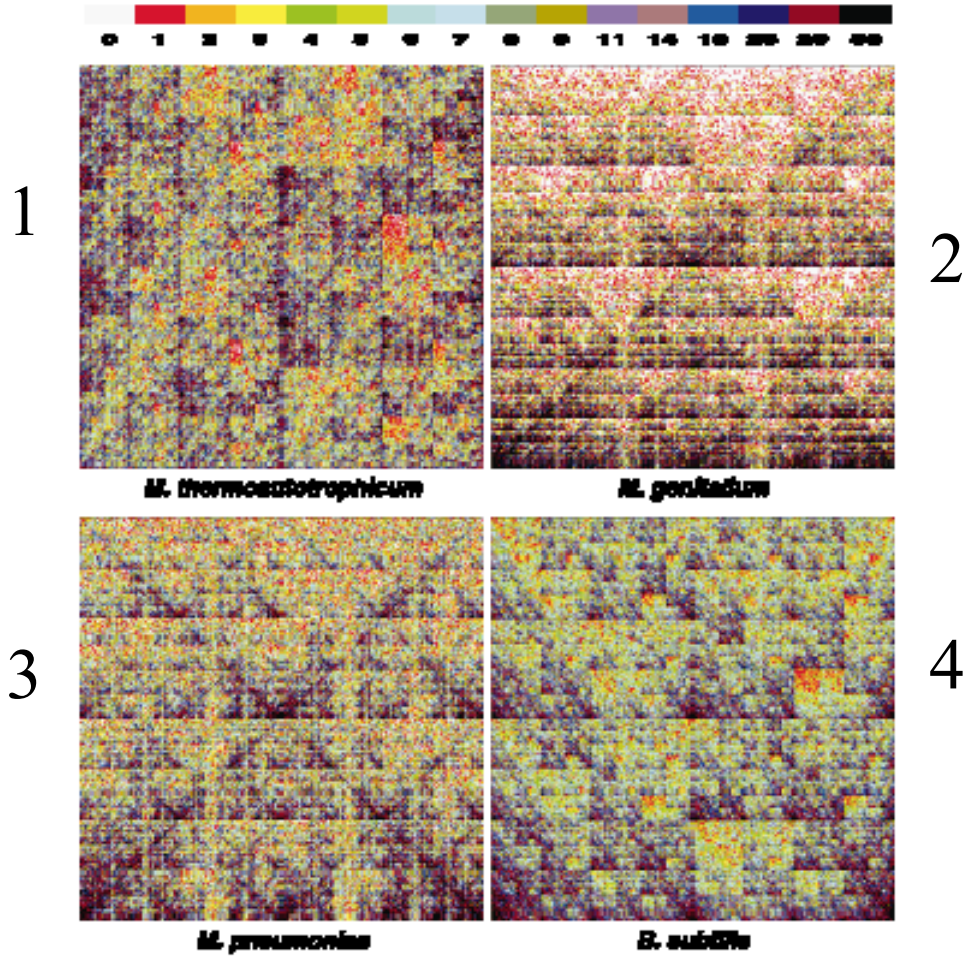


Fig. 4. Portraits of another 4 bacteria in $K = 8$ frames. Note the common crossing patterns in the two Mycoplasma.

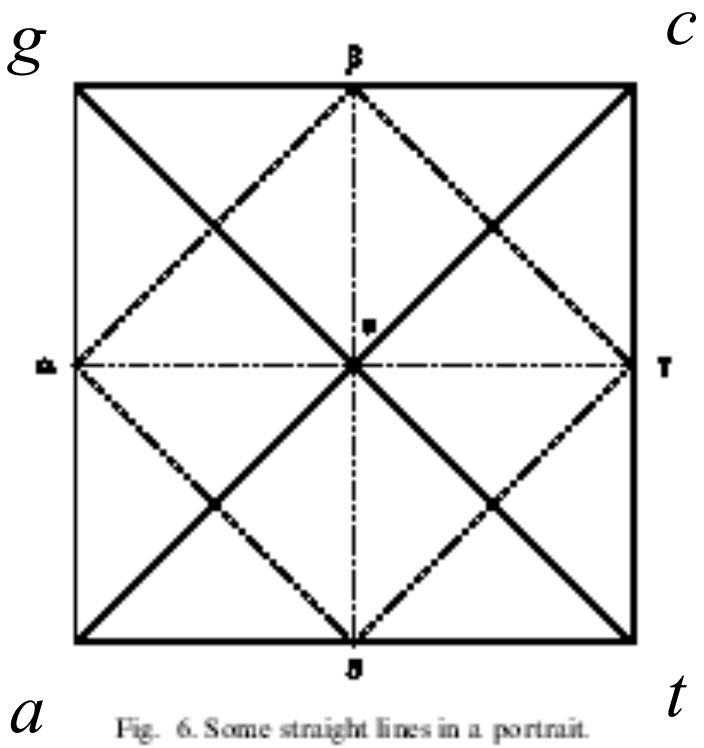
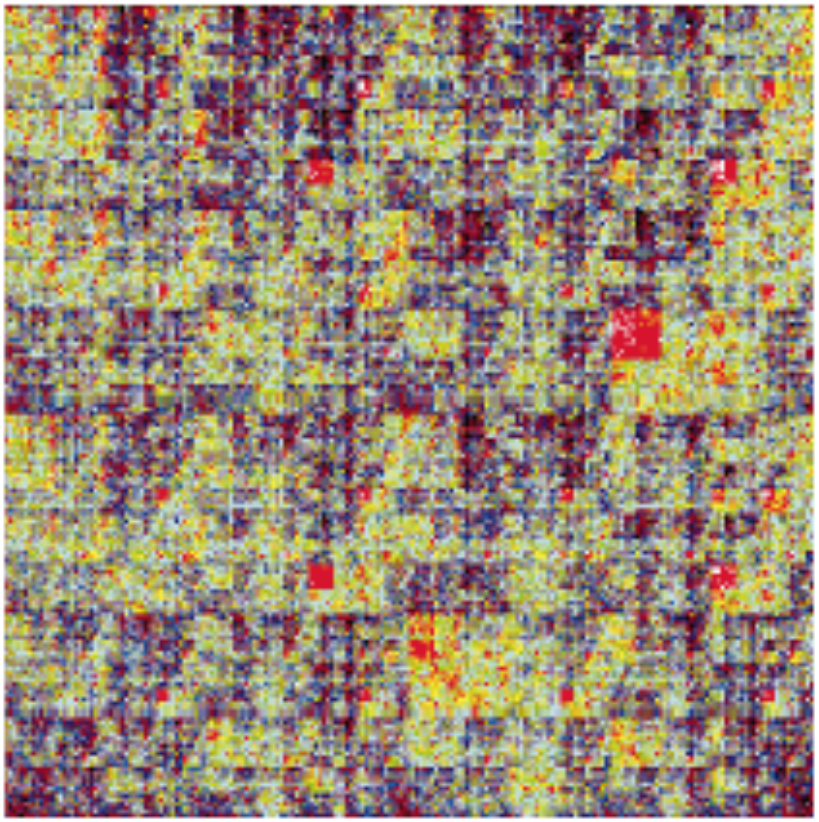


Fig. 6. Some straight lines in a portrait.

"Fractal" (pattern of red squares) caused by extreme under-representation of the palindrome ACGT

$k=8$



E. coli

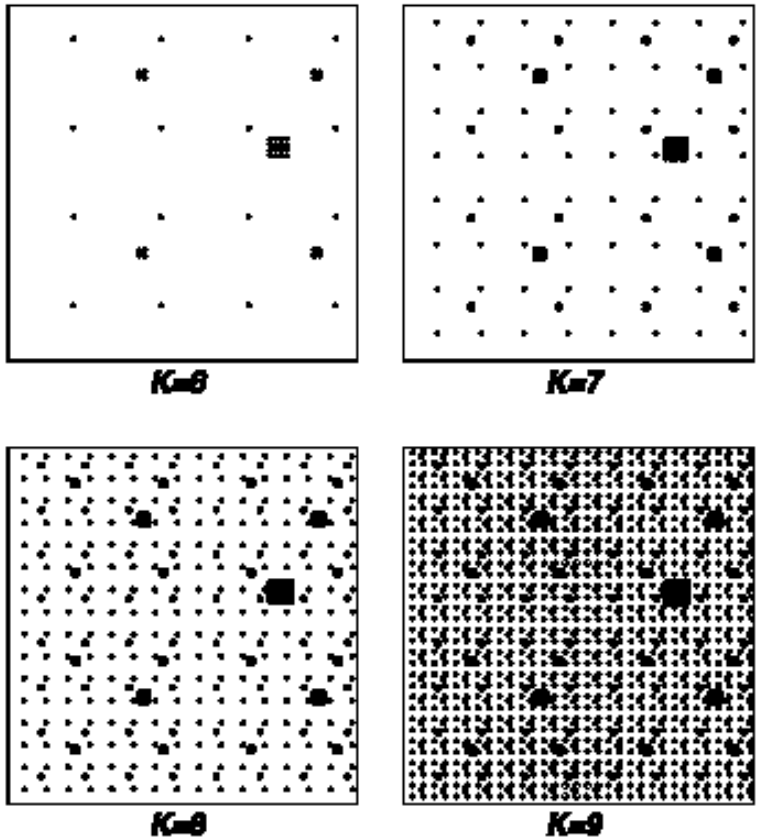


Fig. 7. The location of *cog-tagged strings in K = 6 to 9 frames.*

Frequency set, *k*-spectrum & relative spectral width

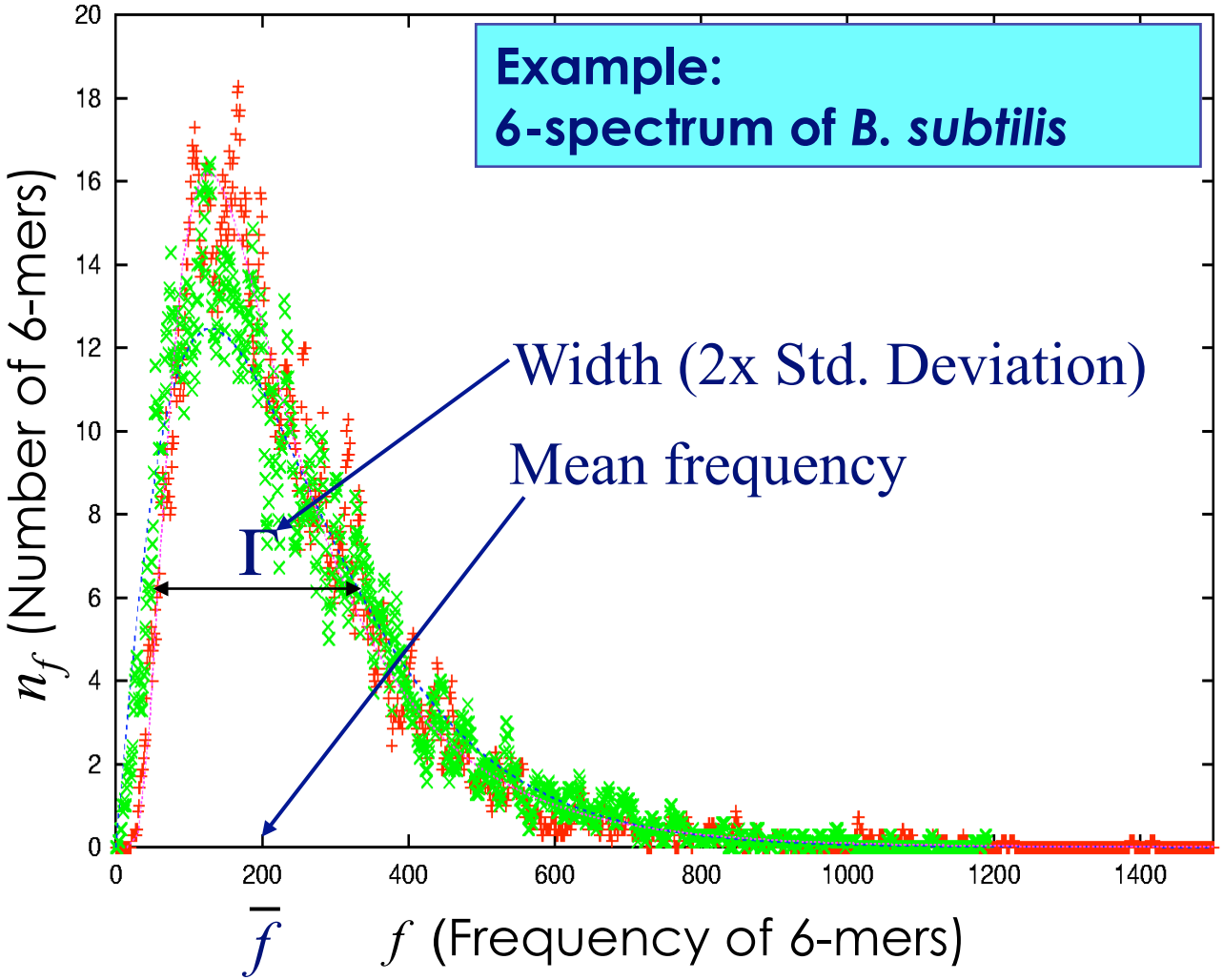
Given freq. set $\{f_i\}$, define

***k*-spectrum**

$\{n_f \mid f=1,2,\ldots\}$

$\Sigma_i f_i = \Sigma_n f\, n_f$

Relative spectral width



Shannon entropy

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

$$H = - \Sigma_i f_i/L \log (f_i/L) = - \Sigma_f n_f \; f/L \log (f/L)$$

- Suppose there are τ types of events: $\Sigma_i = \tau$. 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.

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

Shannon information & relative spectral width

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

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

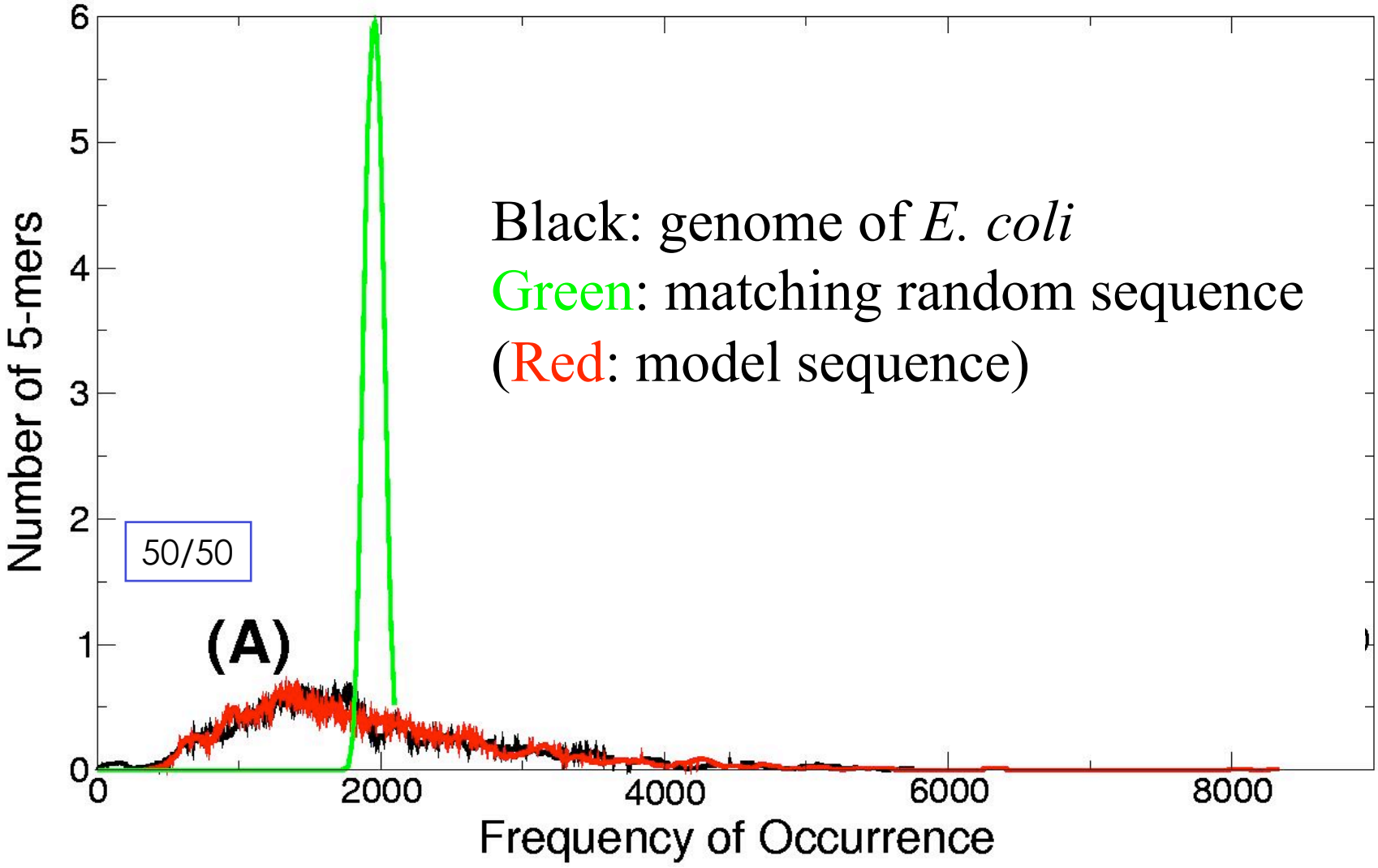
Shannon called R/H_{max} redundancy; Gatlin (1972) called R divergence

- Relation to **relative spectral width** $\sigma = \Gamma/2\,\bar{f}$ (for unimodal distribution)

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

- Shannon information and relative spectral width are equivalent measures

Huge difference between genomes and random sequences



Genomes violently disobey large-systems rule

- Random sequence: width ~ $L^{1/2}$, hence $\sigma \sim 1/L^{1/2} \longrightarrow 0$ for large L
  - *i.e., large systems have sharply defined averages*
- Genomes: $\sigma_{genome} >> \sigma_{random}$
  - *Widths of genomes do not decrease with L*
- Genomes have far more (Shannon) information than random sequences

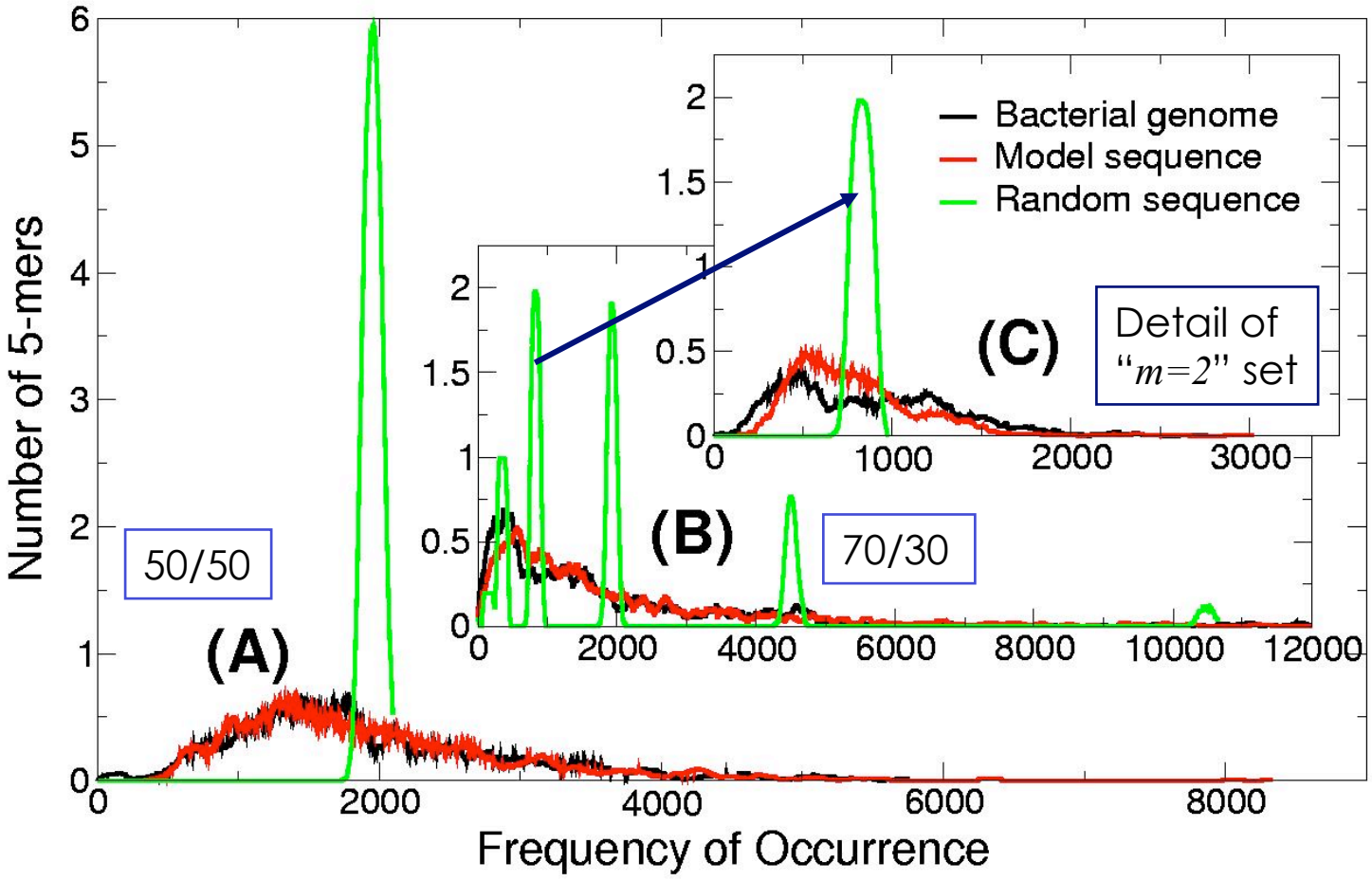
$R = log\,\tau - H$ is a good definition

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**

	Random sequence			Genome sequence		R_{gen}/R_{ran}
k	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	**-**

When A+T ≠ C+G, *k*-spectrum is superposition of *k+1* subspectra

Random sequence: **(A)** Single peak when A+T and C+G same. **(B)** Otherwise split into *k+1* "*m*"peaks, *m*=0 to *k*. Under each *m* peak is spectrum of subset of *k*-mers with *m* A+T's.



(C) Detail of subspectrum of m=2 set. Otherwise split into *k+1* "*m*"peaks, *m*=0 to *k*. Under each *m* peak is spectrum of *k*-mer with *m* A+T's.

Information in 70/30 sequences

Table 2: Shannon information of subspectra $\mathcal{F}_{k,m}$ from the genome *C. muridarum* and corresponding random sequence.

Sequences have AT/CG= 70/30

$k,\ m$	$\bar{f}_m$	R_{Cmur}	R_{random}	R_{ex}	R_{gen}/R_{ran}
2, 1	52,500	7.39 E-3	5.12 E-6	4.76 E-6	1440
3, 2	18,375	2.07 E-2	2.13 E-5	2.04 E-5	963
4, 2	2,756	8.58 E-2	1.75 E-4	1.50 E-4	490
5, 3	964	1.10 E-1	5.10 E-4	4.86 E-4	216
6, 3	145	2.04 E-1	3.42 E-3	3.34 E-3	60
7, 4	50.6	2.61 E-1	9.90 E-3	9.72 E-3	26
8, 4	7.60	4.79 E-1	6.59 E-2	6.53 E-2	7.3
9, 5	2.65	3.05 E-1	1.89 E-1	1.88 E-1	1.6
9, 7	14.5	3.03 E-1	3.43 E-2	3.43 E-2	8.8
10, 6	0.93	1.02 E 0	5.44 E-1	5.37 E-1	1.9
10, 8	5.06	4.24 E-1	0.99 E-1	0.99 E-1	6.2

Reduced spectral width & Shannon information

- Recall k-spectrum superposition of $k+1$ peaks
- For each peak, define

$$\mathbf{M}_\sigma = (\sigma_{genome'}/\sigma_{random})^2$$

and

$$\mathbf{M}_R = R_{genome'}/R_{genome}$$

- For whole k-spectrum, define **reduced spectral width** (**RSW** $\mathbf{M}_\sigma$)and **reduced Shannon information** (**RSI** $\mathbf{M}_R$) **averaged over subspectra**

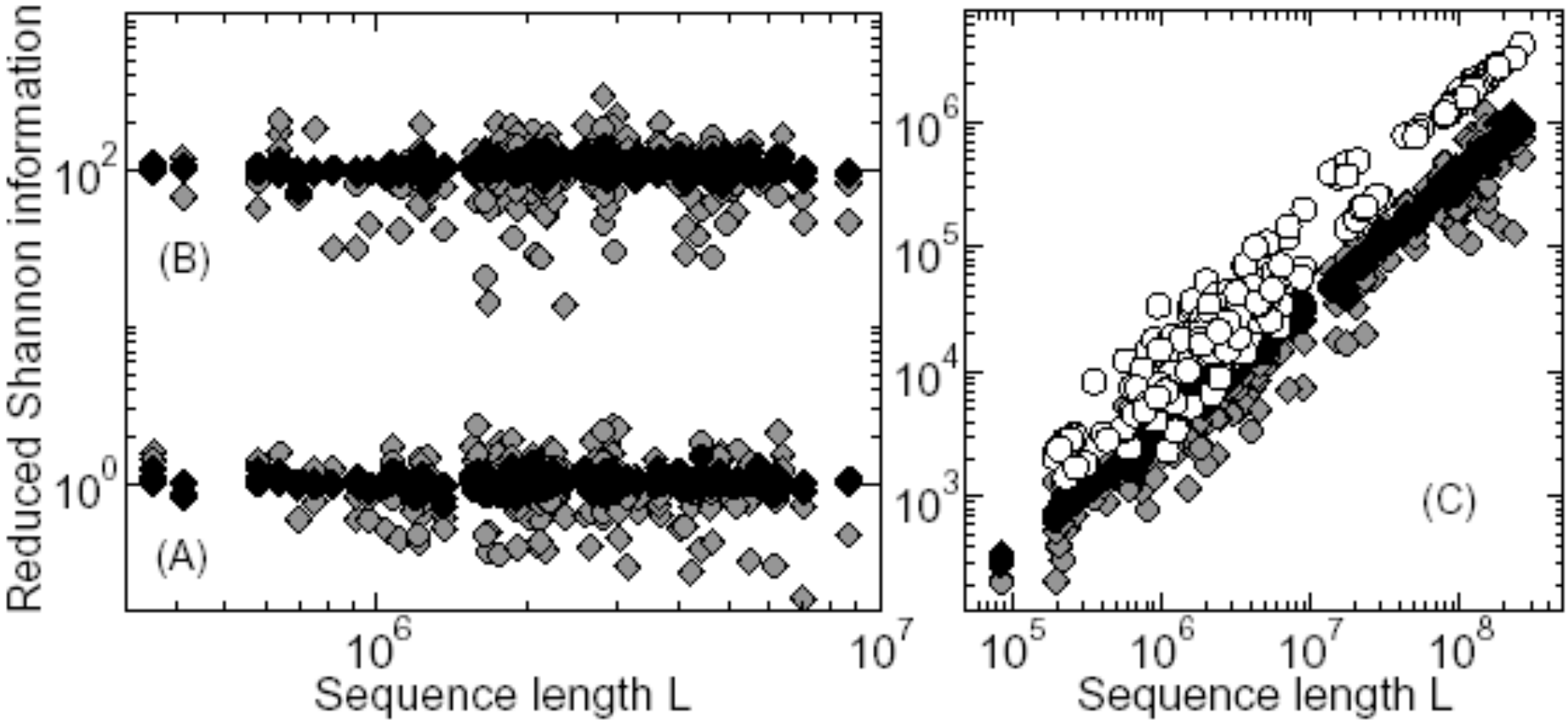
$$\mathbf{M}_\sigma(Q) = \langle\sigma^2/\sigma_{random}{}^2\rangle, \quad \mathbf{M}_R(Q) = \langle R/R_{random}\rangle$$

- Expect

$$\mathbf{M}_\sigma(Q) \sim \mathbf{M}_R(Q), \qquad \mathbf{M}_\sigma(Q_{ran}) \sim \mathbf{M}_R(Q_{ran}) \sim 1$$

Testing M_R (Q_{ran}) ~ 1

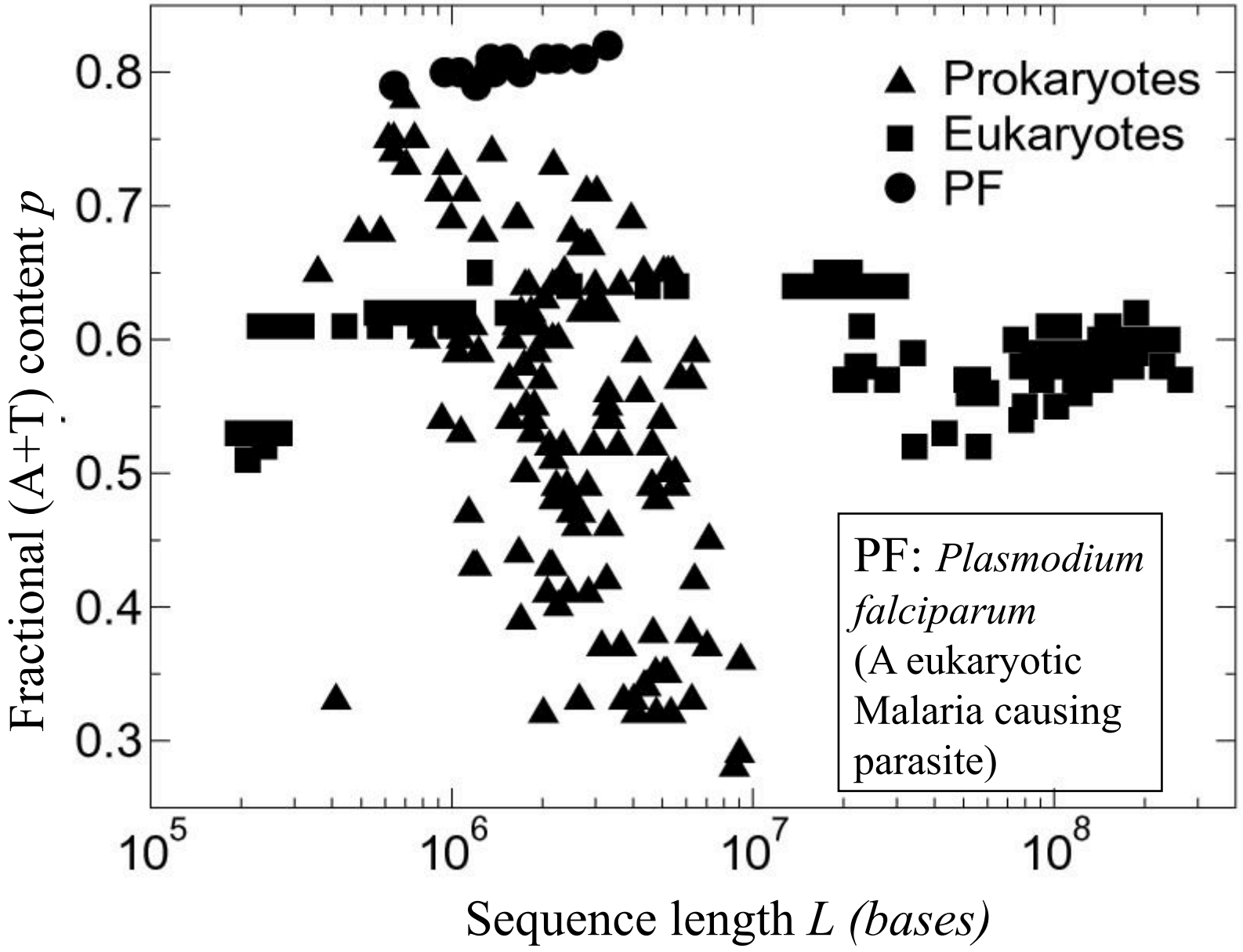
(A) Random "matches" of 155 microbial genomes; *k*=2-10
(B) 100-replica matches of 155 microbial genomes; *k*=2-10



A look at Complete Genomes

~ A universality is discovered

Complete Genomes are diverse

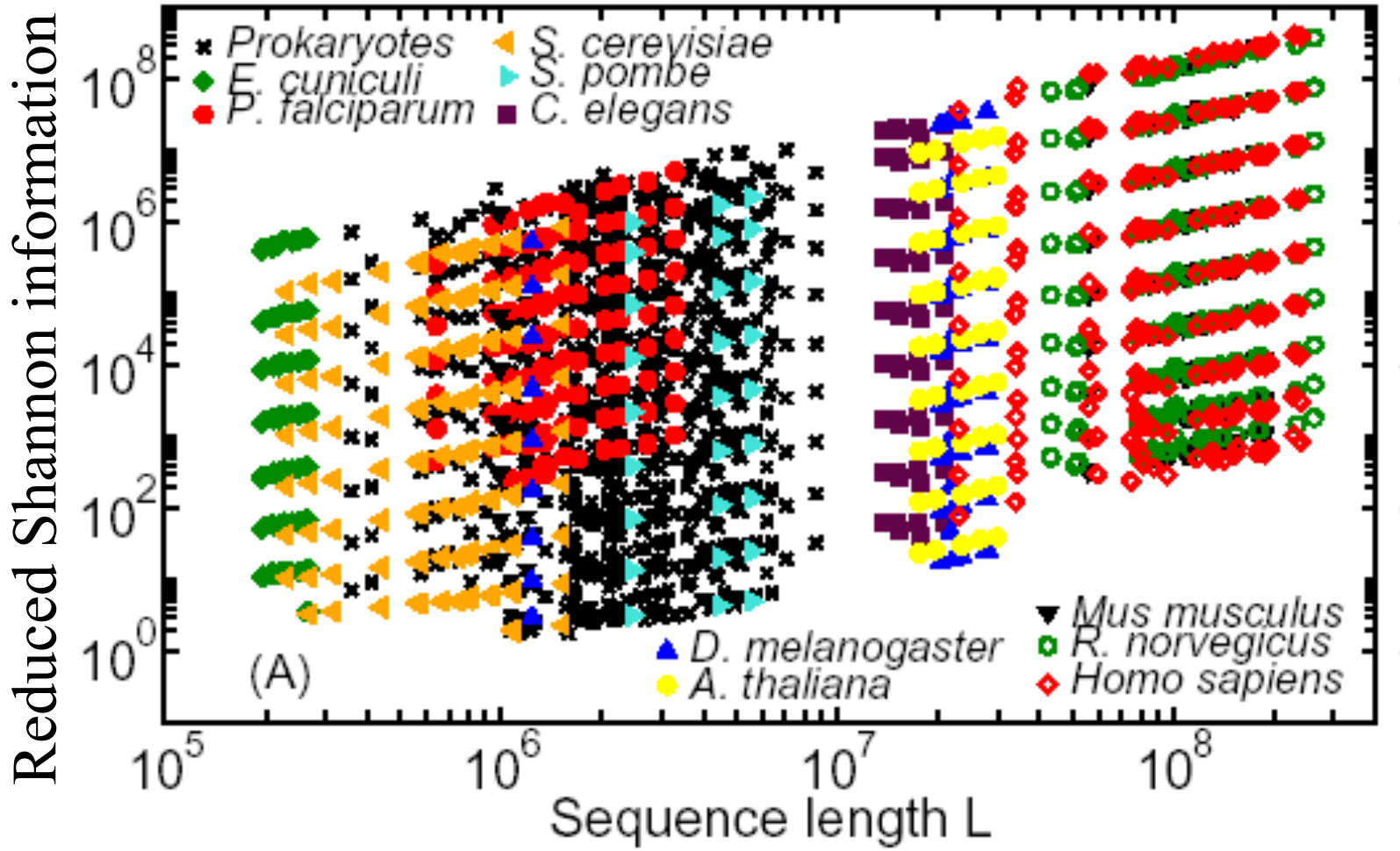


Measurements

- Measure (by computation)
  - reduced spectral widths $\mathbf{M}_\sigma$
  - reduced Shannon information $\mathbf{M}_R$
  - k-spectra, k = 2 to 10
  - 282 complete sequences (155 microbial genomes and 127 eukaryotic chromosomes)

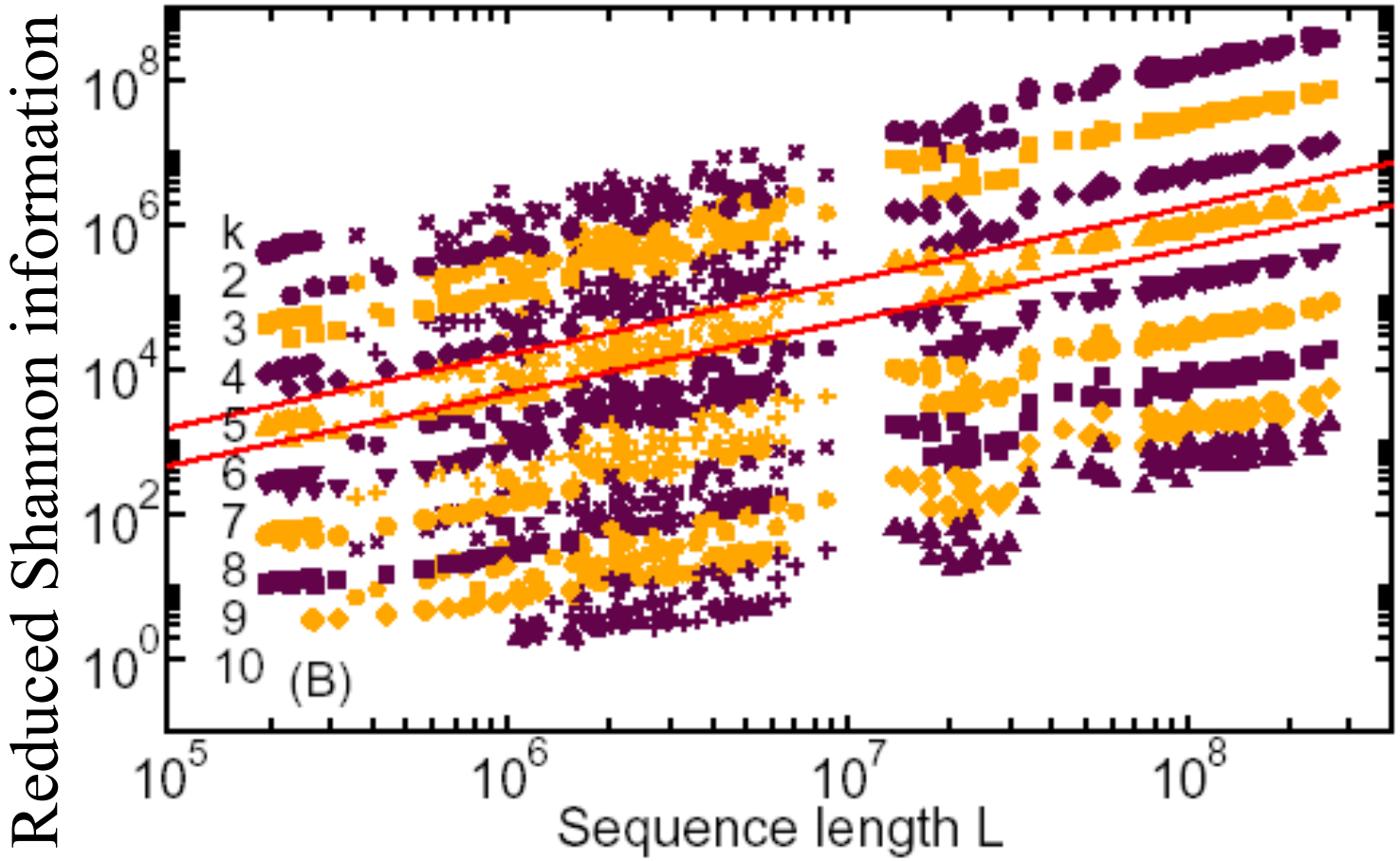
- Results
  - $\mathbf{M}_\sigma \sim \mathbf{M}_R$
  - Plot $\mathbf{M}_\sigma$ versus *L*, sequence length

Results: color coded by organisms



Each point from one k-spectrum of one sequence; >2500 data points. Black crosses are microbials. Data shifted by factor 2^{10-k}.

Color coded by *k*: Narrow *k-bands*



Data from 14 *Plasmodium* chromosomes excluded; ~2400 data points. For each *k*, 268 data points form a narrow $M_\sigma \sim L$ "*k-band*".

k-bands

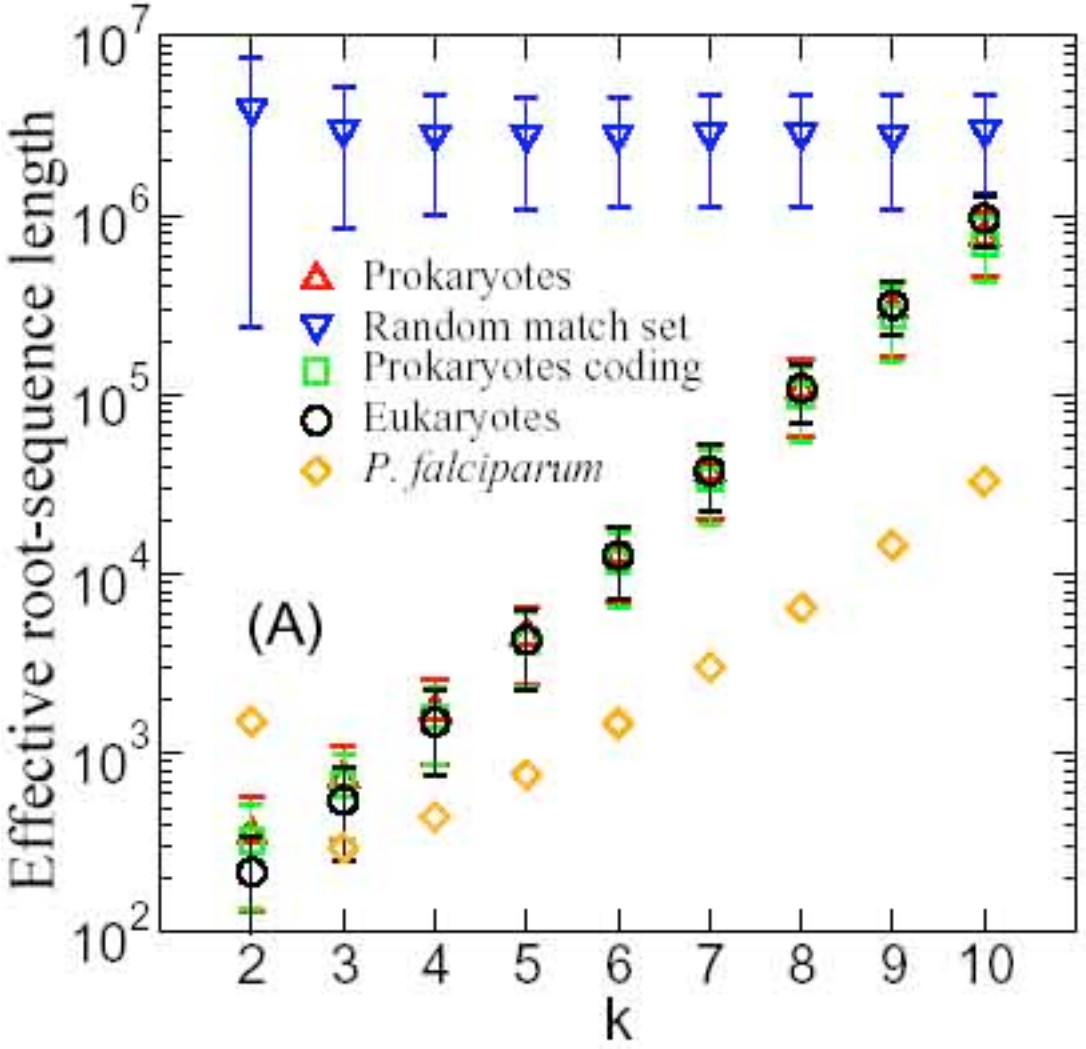
- $\mathbf{M}_R$ is very large
- For each k all data (268 sequences) form a k-band
  - $\mathbf{M}_R/L$ ~ **universal constant** (i.e., same for ALL genomes)

A Universality Class

- Each *k-band* defines a **universal constant** $L/M \sim$ constant $= L_r$ (Effective root-sequence length)
- Obeys

  log L_r(k) = *a k + B*

  1989 pieces of data given be two parameters.
  a = 0.398+-0.038
  B = 1.61+- 0.11
- Defines a **universal class**
- Plasmodium has separate class:
  a = 0.146+-0.012



Black: genome data; green: artificial

Replicas & Root-Sequence Length

~ How to create information stochastically

Replica & universal root-sequence length

- Take random root-sequence of length L_r and replicate to length L of some genome, then full sequence will have

  $$\boldsymbol{M}_R = L/L_r \qquad \text{(for any } k\text{)}$$

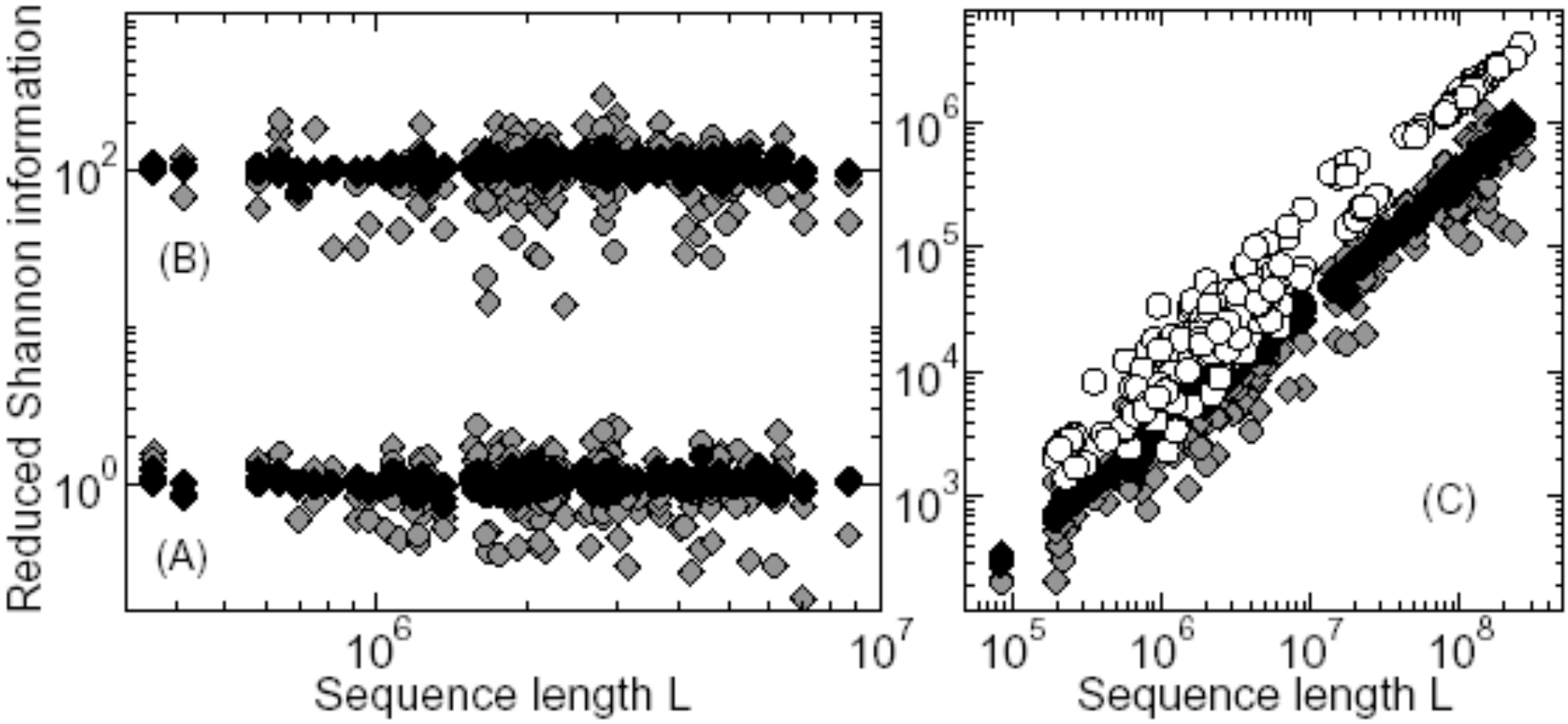
- Or, any sequences obtained by replication of the root-sequence (*i.e.* a *replica*) will have

  $$L/\boldsymbol{M}_R = L_r$$

- A set of replicas of variable lengths all replicated from (not necessarily the same) random root-sequences of length L_r will have *k*-independent universal $L/\boldsymbol{M}_R = L_r$

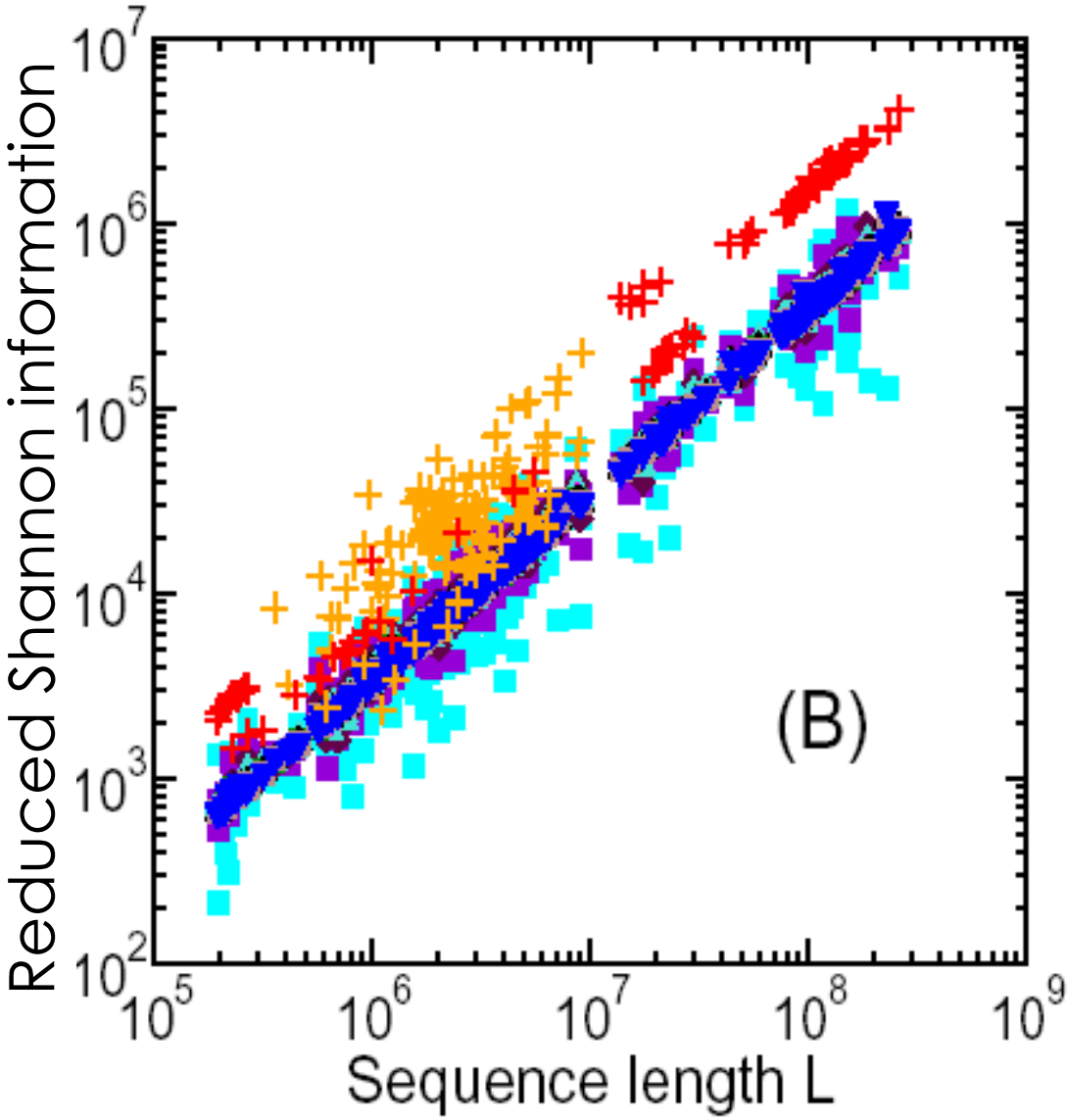
RSI in an m-replica is multiplied m times

(A) Random "matches" of 155 microbial genomes; *k*=2-10
(B) 100-replica matches of 155 microbial genomes; *k*=2-10



Reduced Shannon information In Replicas

- Squares: $\boldsymbol{M}_R$ in *m*-replicas
  - root-sequence length 300
  - 260 replicas match profiles of genomes
  - sky: k=2,
  - purple: k=3
  - blue: k=4-10
- Crosses: $\boldsymbol{M}_R(k=2)$ in genomes
- Replicas like genomes, but lack *k*-dependence



A Model for Genome Growth & Evolution

~ How did life create information stochastically

A Hypothesis for Genome Growth

- Random early growth
  - Random b/c has no information
- Followed by

1. random segmental duplication and
2. random mutation

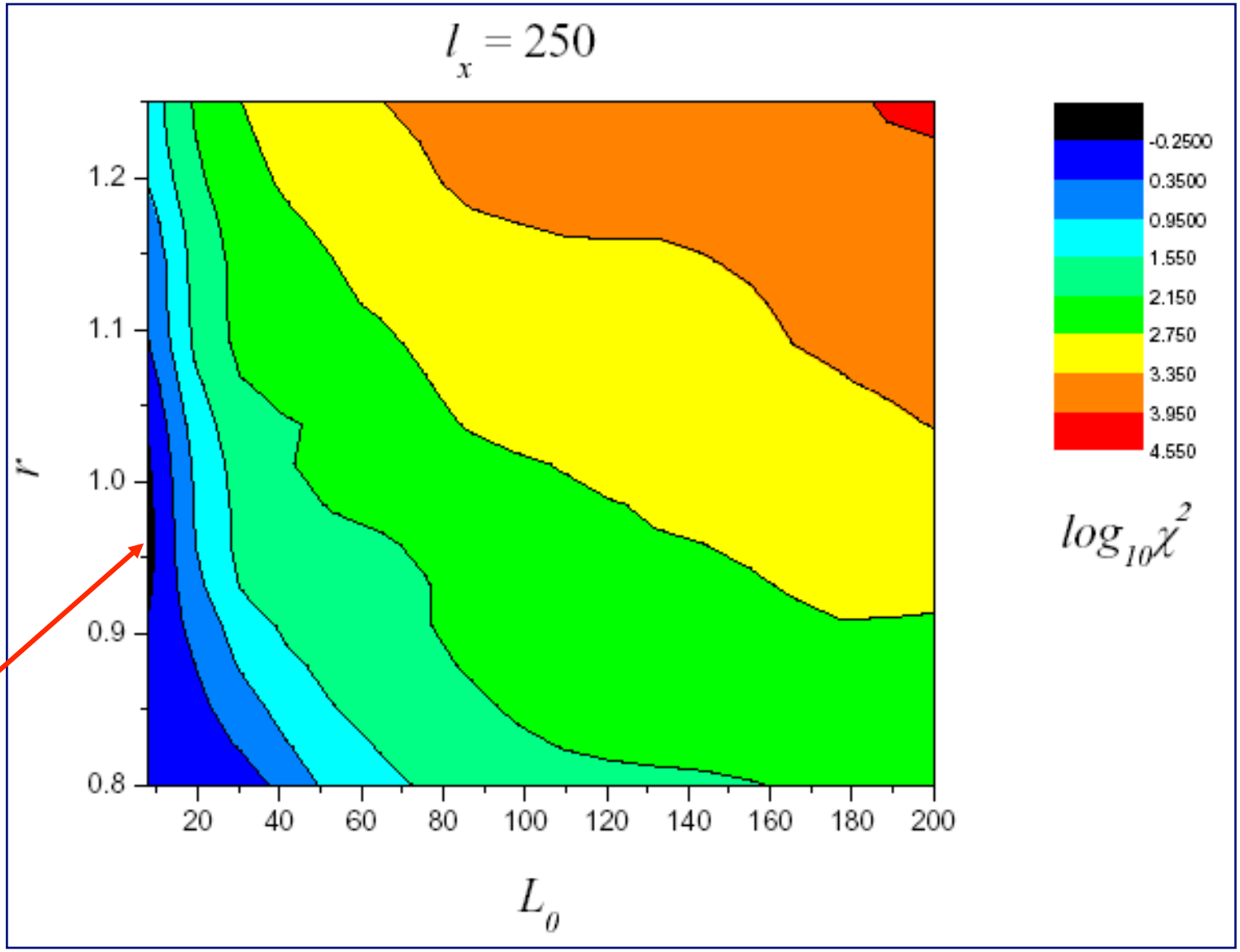
Self copying – strategy for retaining and multiple usage of hard-to-come-by coded sequences (i.e. genes)

The Minimal Model

- Start with length L_0
- Segmental duplication is **maximally stochastic** and grow to full length *L*
  - random selection of site of copied segment
  - weighed random selection $g(l)$ of length of copied segment
  - random selection of insertion site of copied segment
  - Biologists: **replicative translocation**
- Mutation is standard single-point replacement (no insertion and deletion)
  - Point mutation at rate of ***r*** per base

$$\chi^2 = \langle [((L_r)_{model} - (L_r)_{gen})/\Delta (L_r)_{gen}]^2 \rangle$$

Model param-eter search: favors very small L_0



The Minimal Model (cont'd)

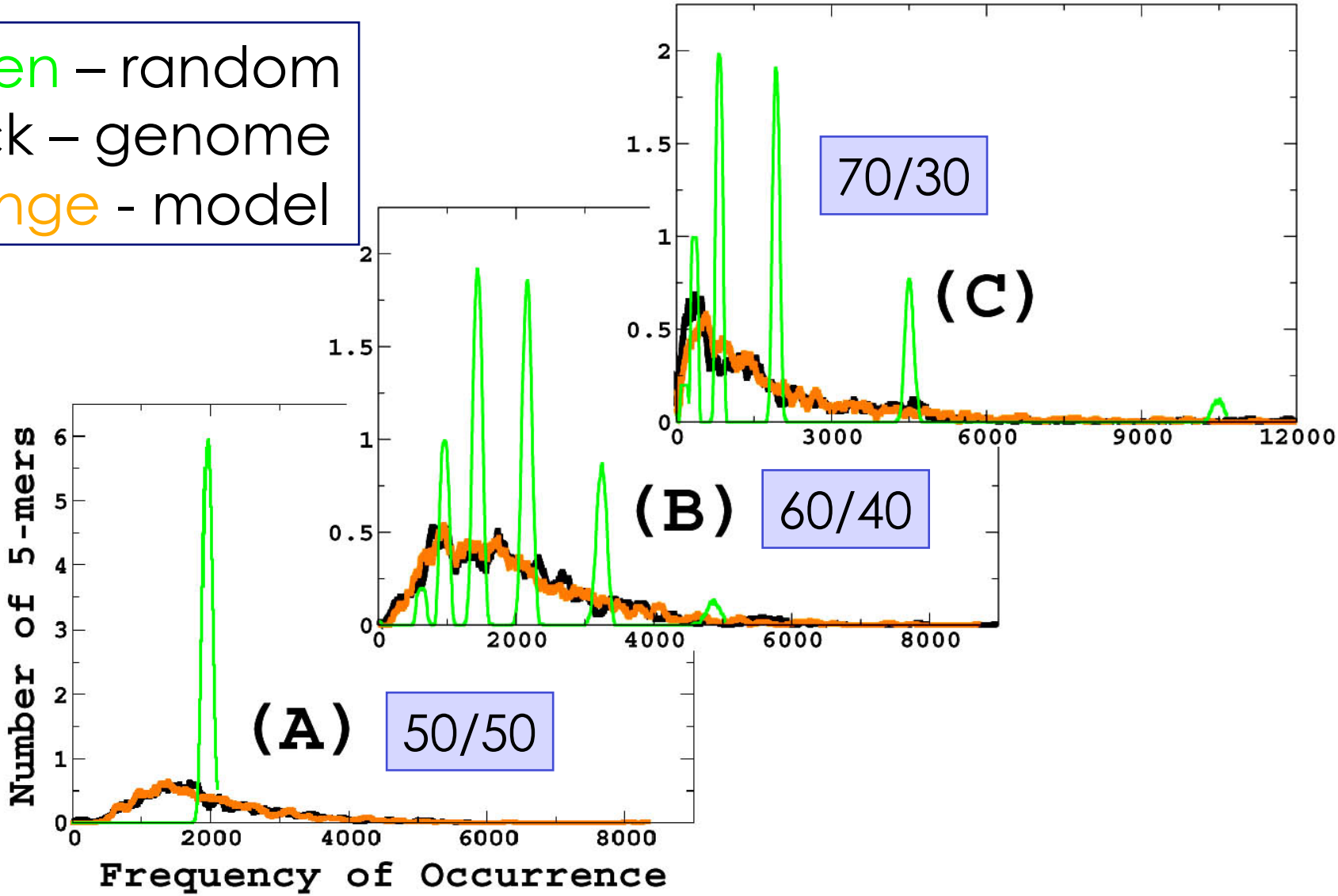
- Best parameters (preliminary; after non-exhaustive search)
  - L_0 ~ 8 *b*
  - ***r*** ~ 0.95~1.1 (mutations per 100 *b*)
  - ***g(l)***: equal probability $0 < l < l_x$

    l_x = 250~2000 if current seq. length L_c< 2 *Mb*

    l_x = *10000 if* L_c> 2 *Mb*
- Generated model sequence set with same length and composition profile as complete genome set
- Computed *k*-spectra, $\mathbf{M}_\sigma$, $\mathbf{M}_\mathbf{R}$, L_r, etc.

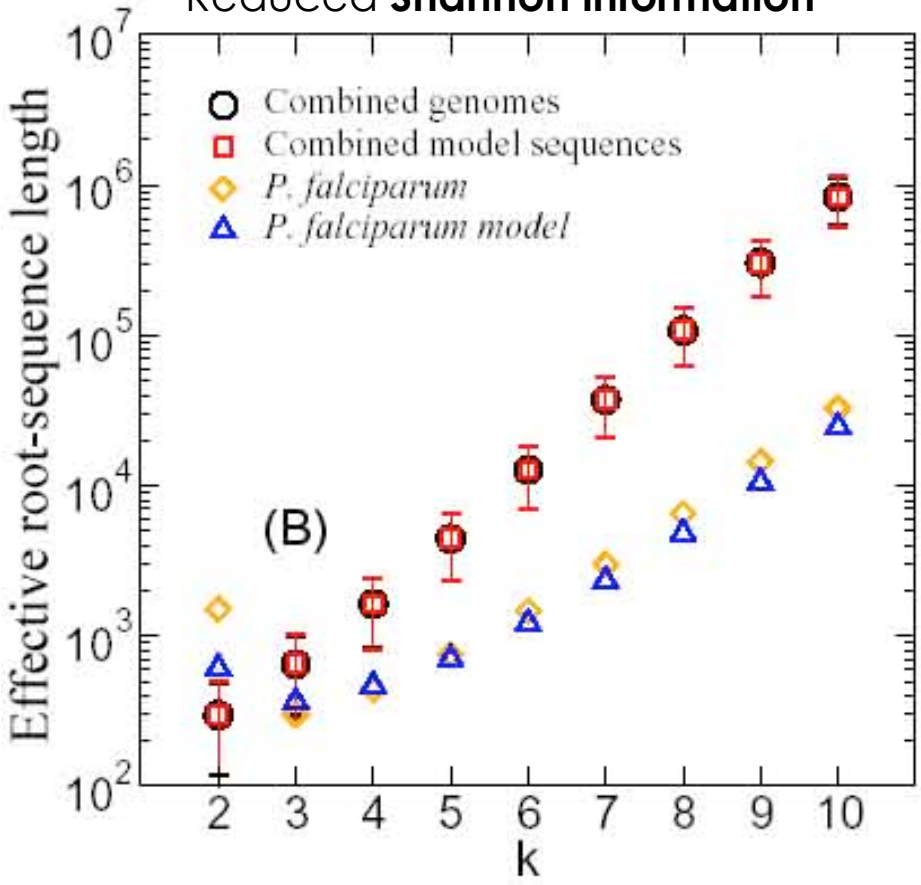
5-spectra of "genomes" with different base compositions

Green – random
Black – genome
Orange - model

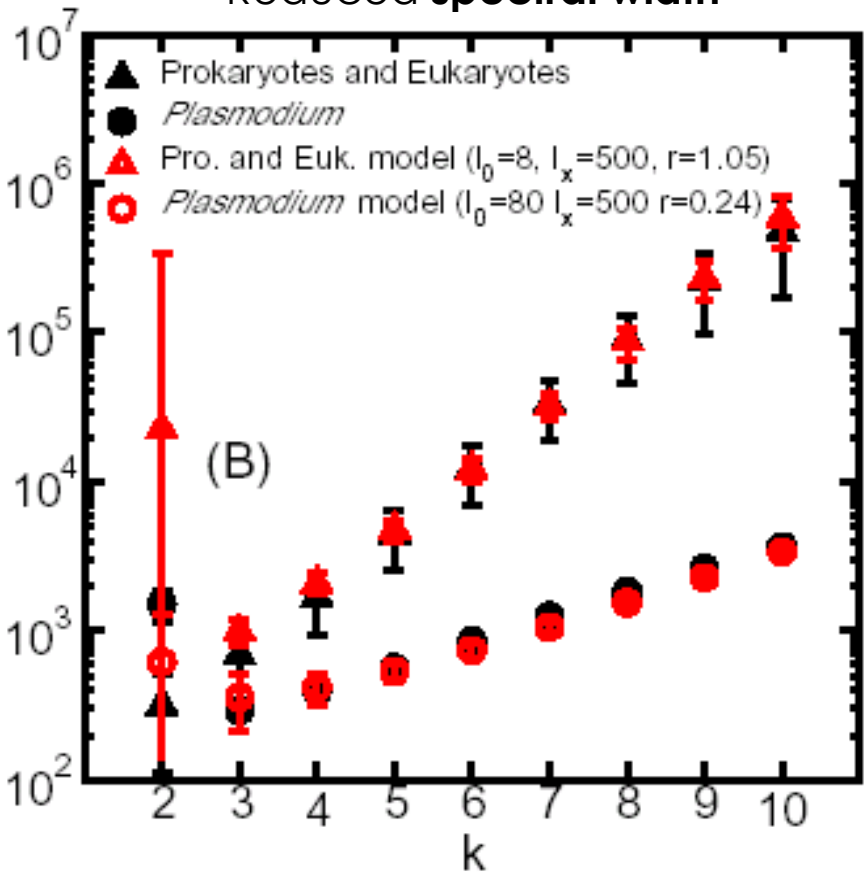


Universality classes

Reduced **Shannon information**



Reduced **spectral width**



Red & blue symbols are from (same) model sequences

Self-Similarity in Genomes

~ Genomes emulates self-organized critical systems

Are genomes self similar?

- Very small L_{eff} suggests genomes has very high duplication content
- Our model based on maximally stochastic segmental duplication reproduce empirical k-spectra and L_{eff}
- If genomes are sufficiently uniform, then genome should exhibit whole-genome property on a scale of ~L_{eff}
  - *i.e.* for any segment of length l, should have

$$\mathbf{M}_{\sigma}(k)/l \sim (\text{RSW of whole genome})/L \sim L_{eff}(k)$$

$M_R(k)$ in 8 randomly selected segments of length $l = L/2^n$

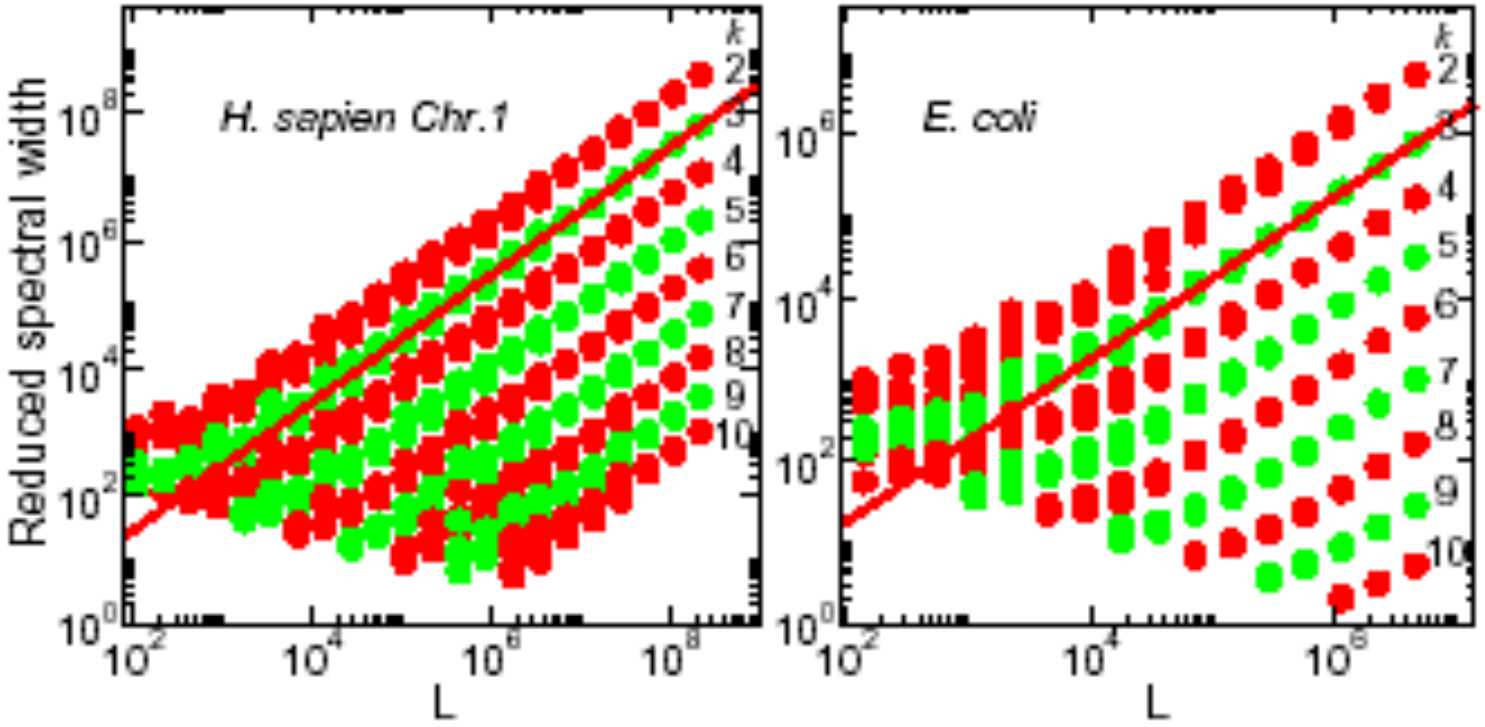


Figure 1: RSW ($\mathcal{M}_\sigma$) of k-spectra, k=2 to 10, of segments from the 246 Mb chromosome I of *H. sapiens*. Lengths of the segments are $1/2^n$ of full length, n=1 to 21, and for each length eight segments are randomly selected. Data for which segment length is less than 4^k are not included. Data for the same k forms a *k-band* approximately linear in L (red line), and each data point has been multiplied by factor of 2^{10-k} to delineate the k-bands for better viewing.

- Given genome length *L* and RSW $\mathbf{M}_\sigma$
- Randomly select set of 25 segments of length l labeled i and compute $\mathbf{M}_{\sigma i}$ of segments
- Define

$$\chi^2(l) = \frac{1}{25}\sum_{i=1}^{n}\left(\left(\log\frac{LM_{\sigma i}}{lM_\sigma}\right)/\log 2\right)^2$$

- If $\chi^2 < 1$ then on average $\mathbf{M}_{\sigma i}/l$ *within factor of 2 of* $\mathbf{M}_\sigma/L$
- Find
  - $\mathbf{L_u}$: segment length above which all sets have $\chi^2 < 1$
  - $\mathbf{L_d}$: segment length below which all sets have $\chi^2 > 1$

L_u and L_d, k=5, all complete sequences

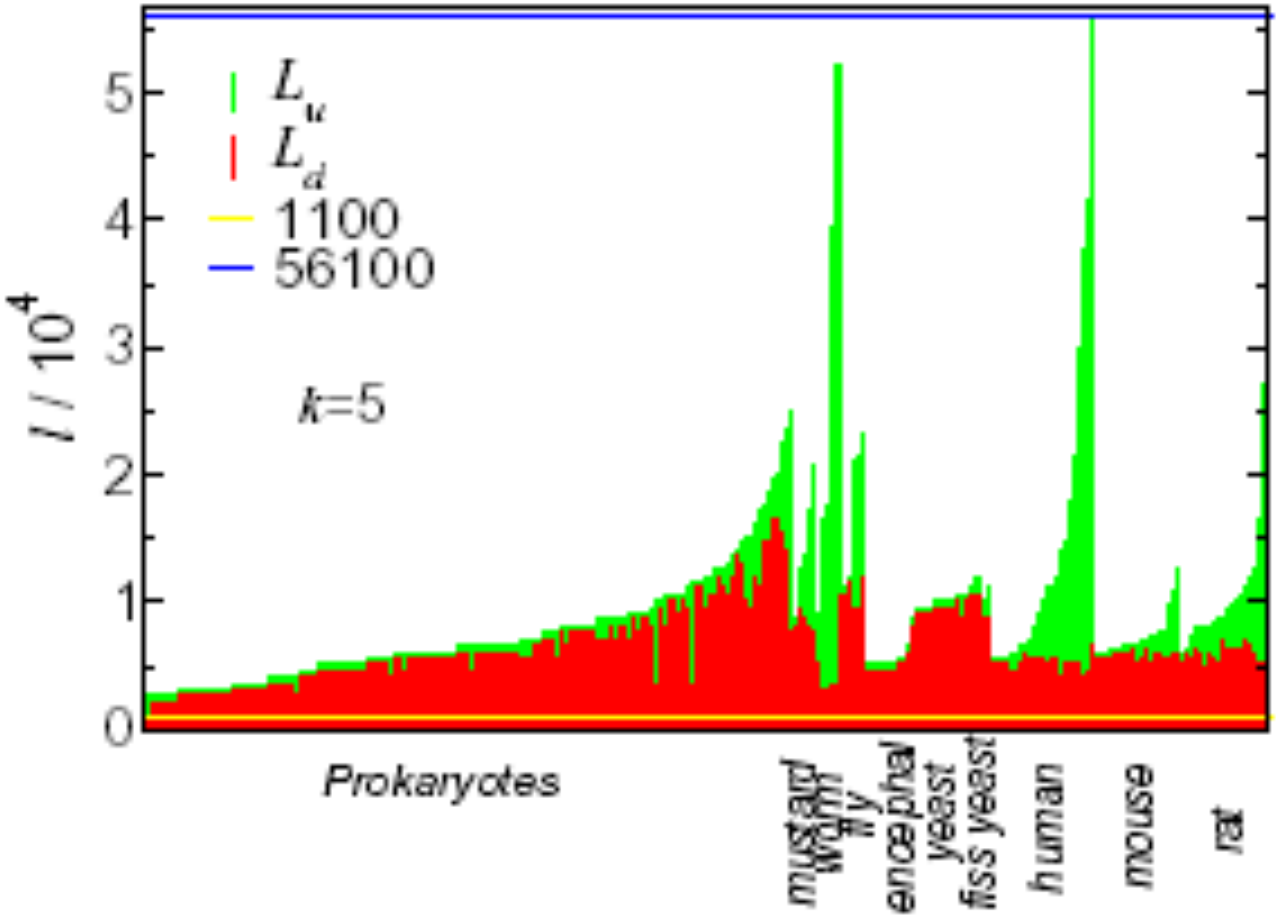


Figure 3: L_u (the length above which all segments are similar to the genome; green bars) and L_d (the length below which no segment is similar to the genome; red bars) for k=5 for all complete sequences in the main universality class. The blue (yellow) line is the position of L_{max} (L_{min}).

L_u and L_d, k=2, 4, 6, 8

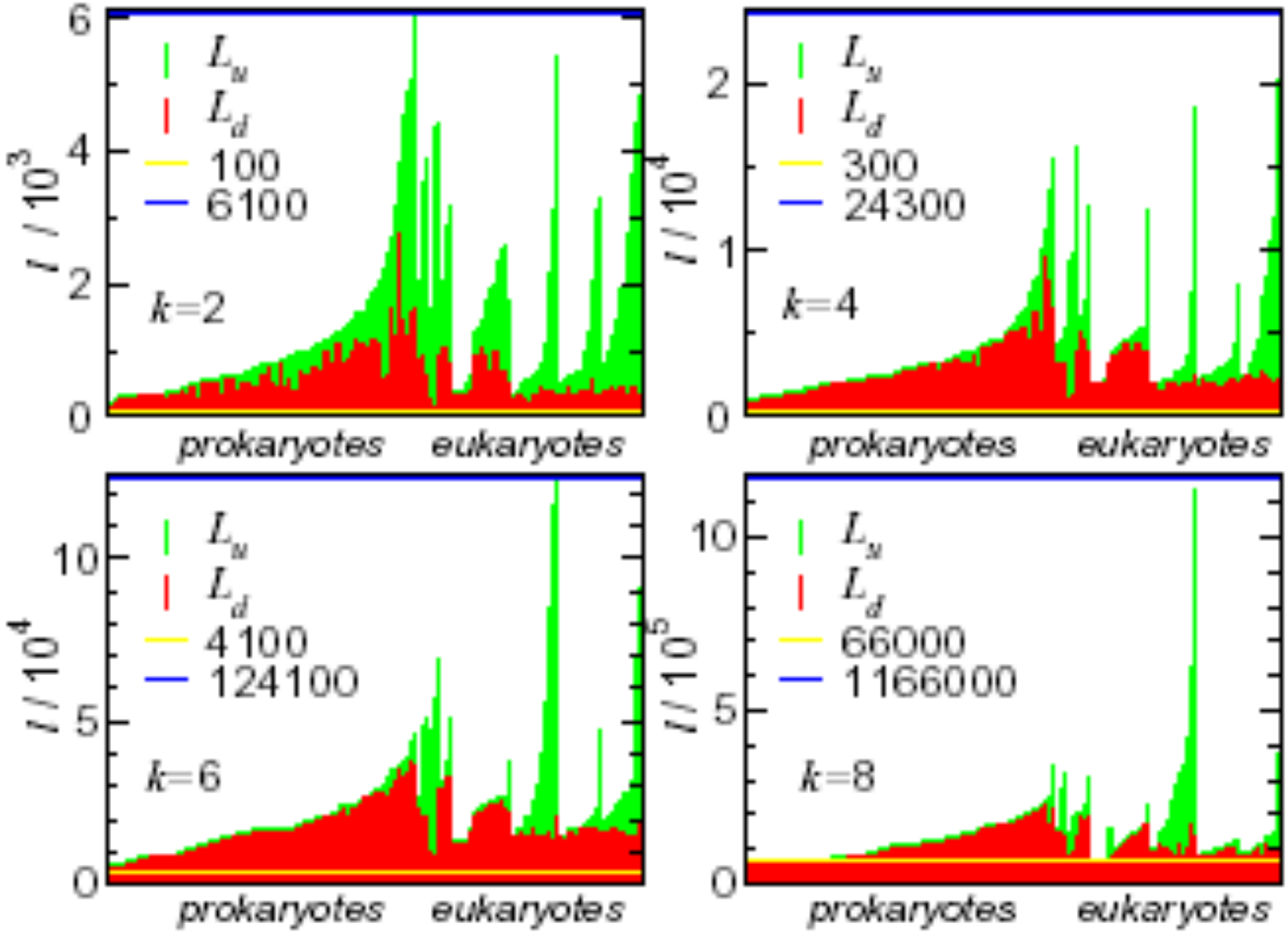


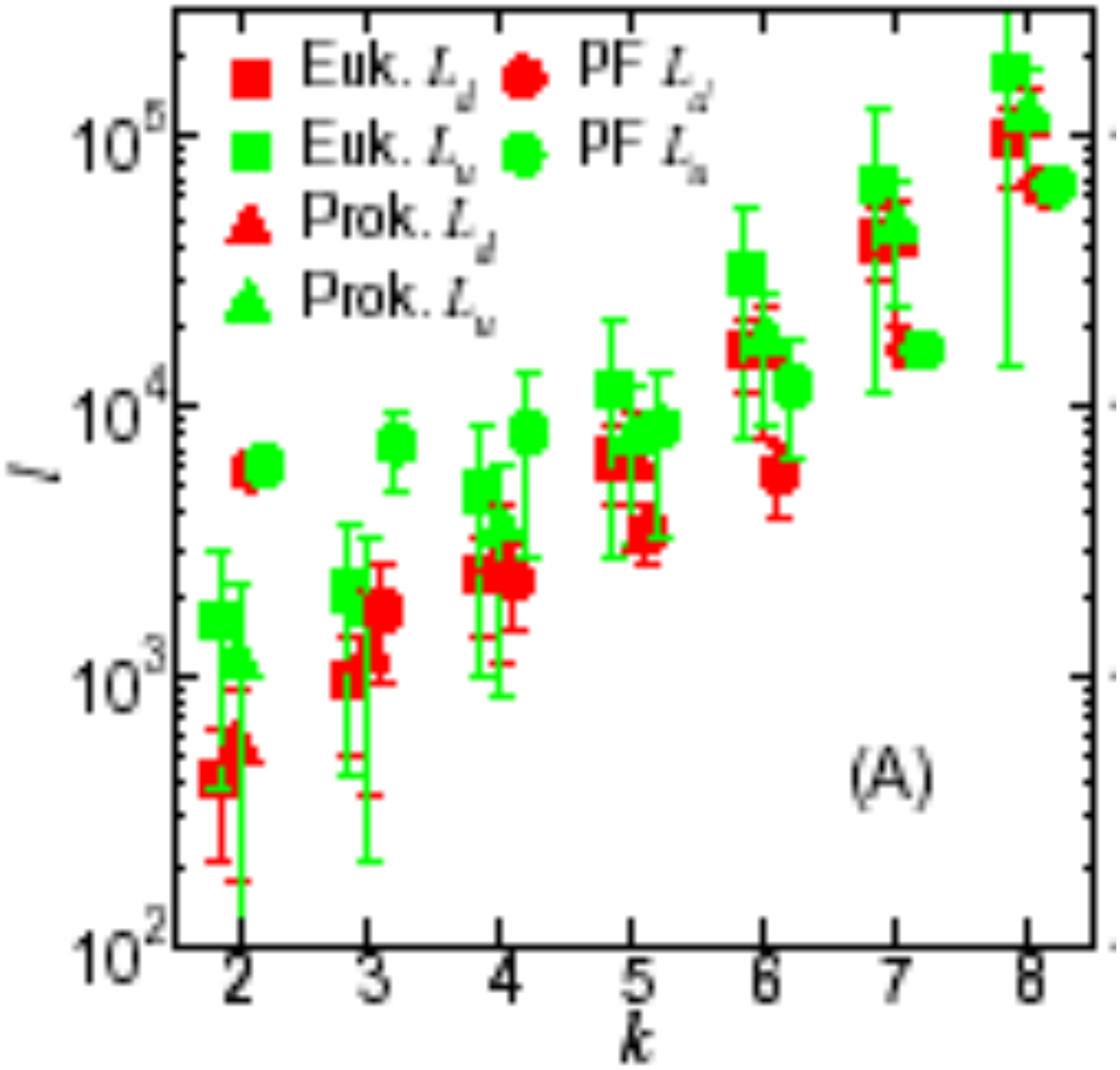
Figure 4: L_u (green bars) and L_d (red bars) for k=2, 4, 6 and 8; see caption for Fig. 3 for more detailed description.

Average Results

	Eukaryotes		Prokaryotes	
k	L_d	L_u	L_d	L_u
2	4.20±2.18 E2	1.63±1.26 E3	5.43±3.69 E2	1.12±1.12 E3
3	9.62±4.43 E2	2.08±1.65 E3	1.20±0.84 E3	1.70±1.50 E3
4	2.35±0.92 E3	4.88±3.89 E3	2.66±1.55 E3	3.47±2.61 E3
5	6.39±2.03 E3	1.18±0.91 E4	6.15±3.17 E3	7.54±4.39 E3
6	1.63±0.48 E4	3.11±2.36 E4	1.53±0.77 E4	1.78±0.92 E4
7	4.12±1.24 E4	6.82±5.71 E4	3.99±1.96 E4	4.60±2.28 E4
8	9.77±3.23 E4	1.76±1.62 E5	1.10±0.45 E5	1.20±0.55 E5

- Prokaryotes: $L_u \sim L_d$
- Prokaryotes $L_d \sim$ Eukaryotes L_d
- Eukaryotes: L_u significantly $> L_d$

Average L_u and L_d versus k



Compare L_{eff} (L_r) with similarity length

Table 3: Comparison of 4^k and mean values of $L_r(k)$ and $L_{sim}(k)$.

k	4^k	$\langle L_r \rangle$	$\langle L_{sim} \rangle$
2	16	310±200	690±570
3	64	680±350	1300±990
4	256	1690±760	2820±1700
5	1024	4450±1900	6690±3200
6	4096	12300±5200	16400±7200
7	16384	33600±15000	42700±18000
8	65536	89500±43000	109000±44000

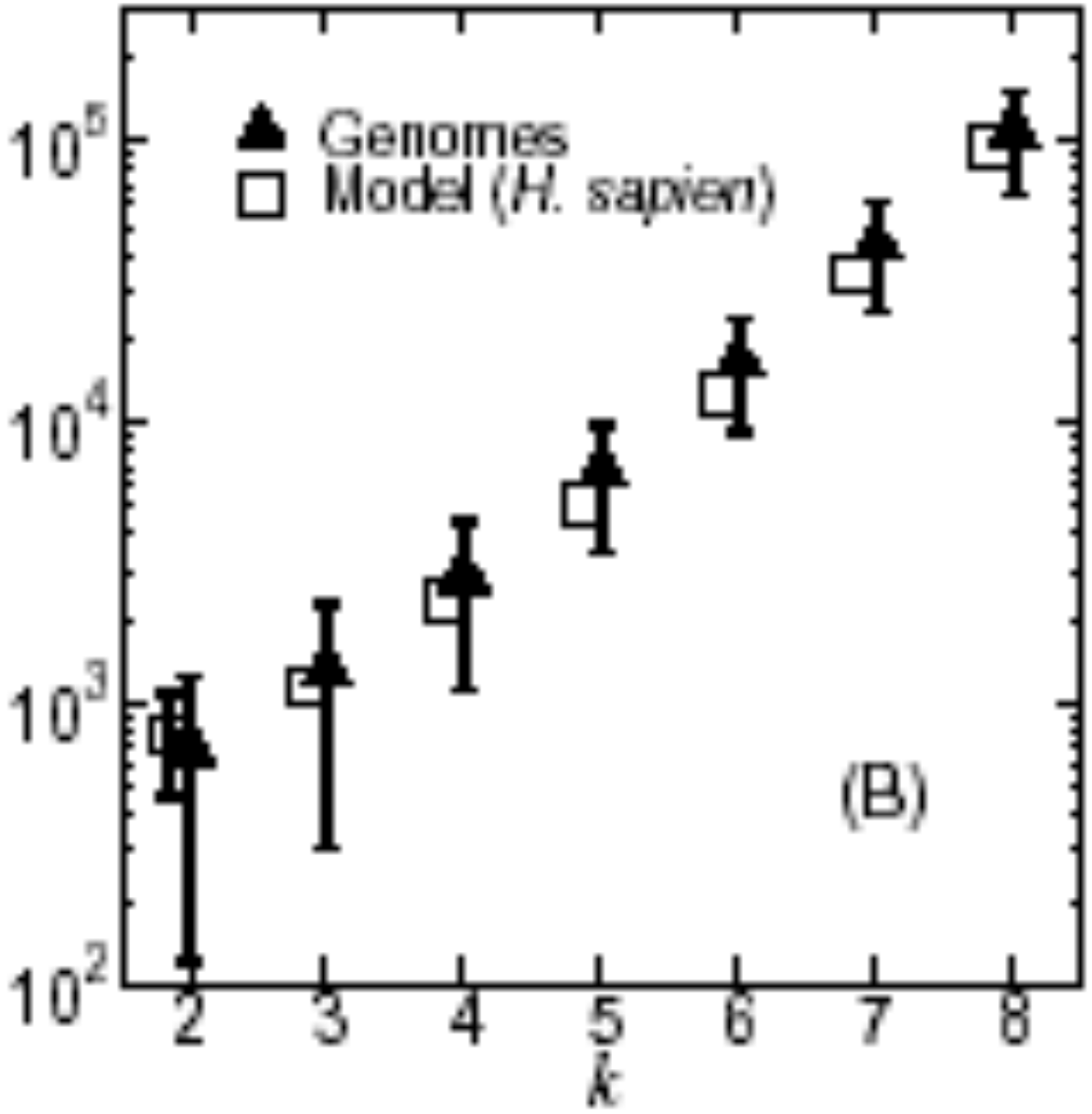
- L_{sim} is the average of prokaryotic L_u and L_d eukaryotic L_d
- L_{sim} barely L_r > barely > 4^k,
- Hence **genomes are almost maximally self-similar**

Compare genomic and model L_{sim}

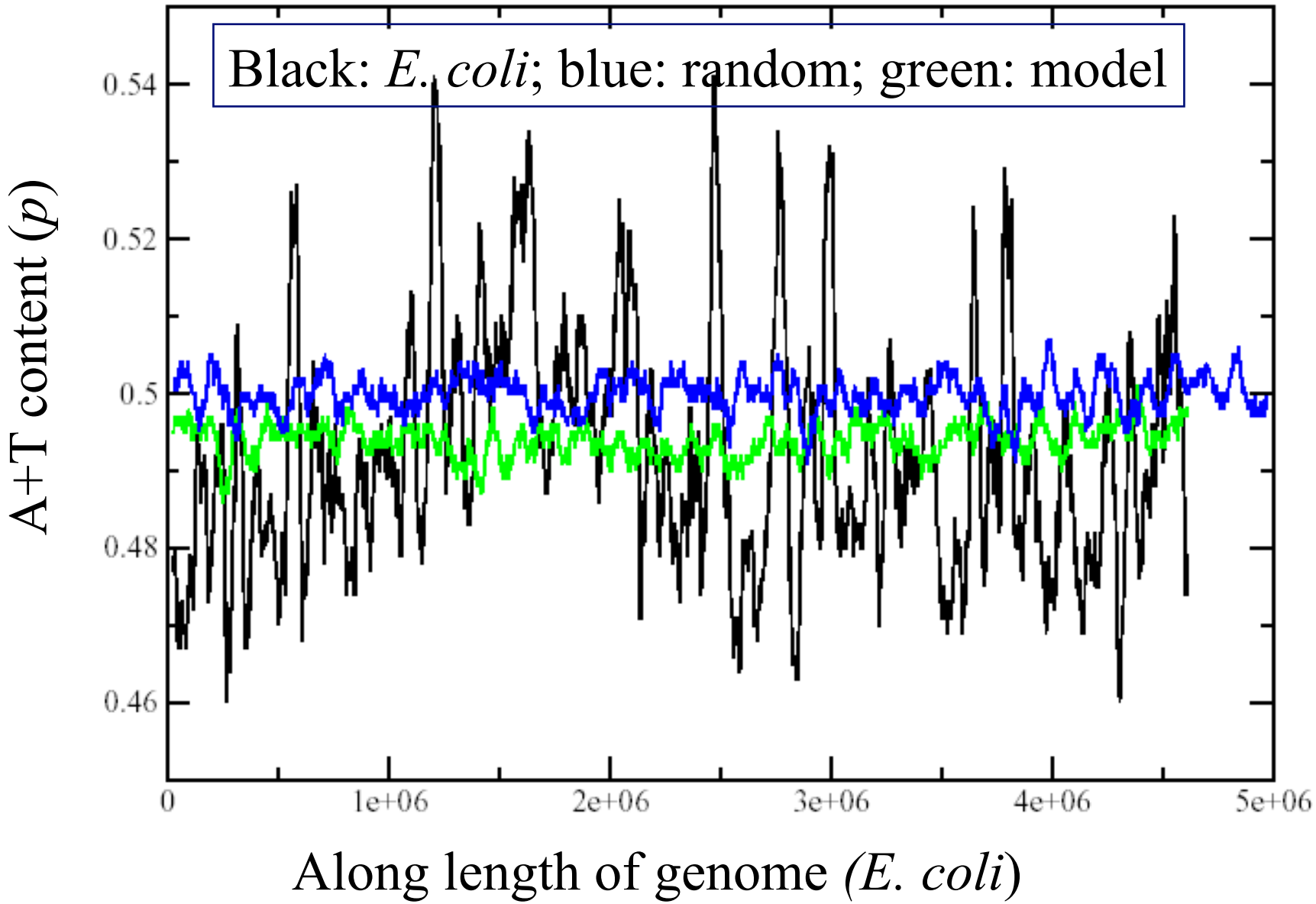
Note: Model predates data

But model has **smaller spread**

Model is too smooth



Texture of genome are rougher then model

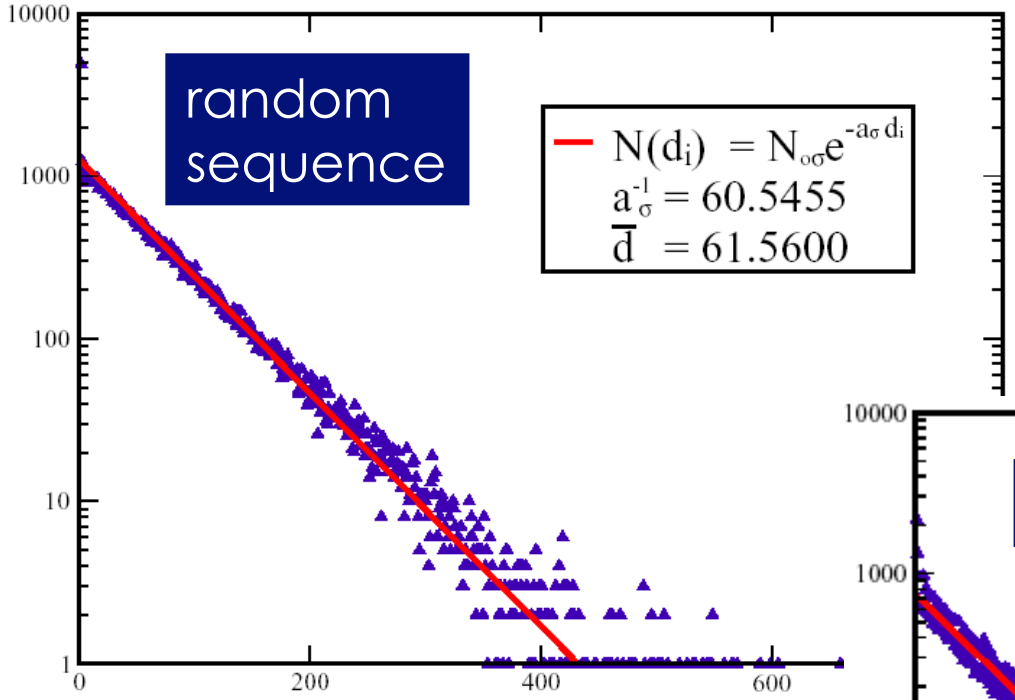


Randomness in Genomes

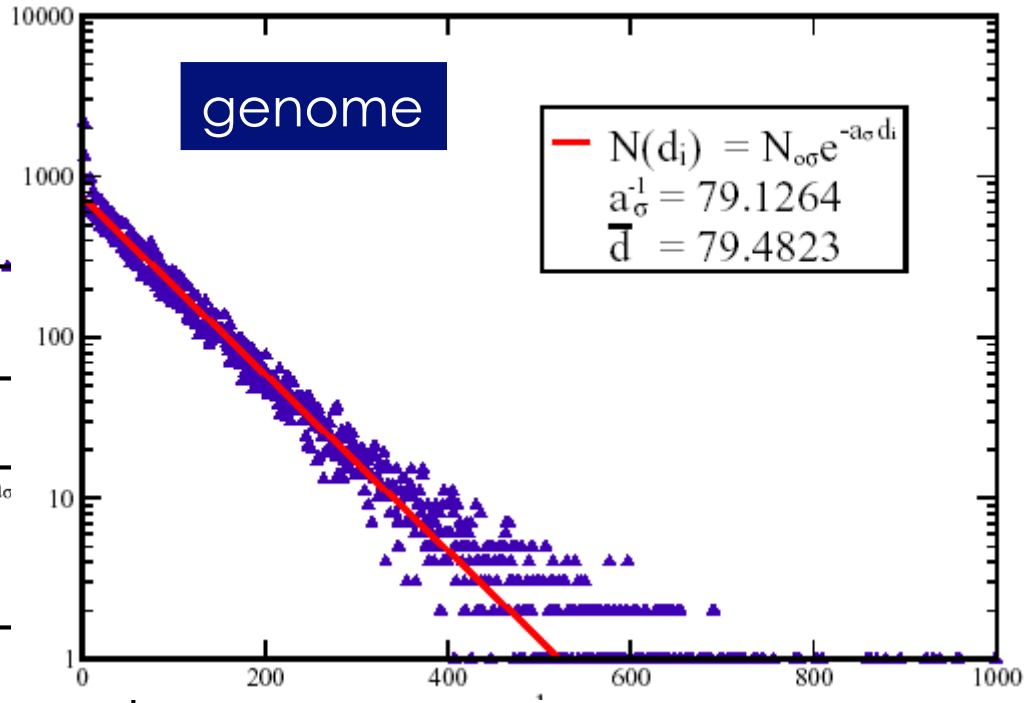
~ Genomes are not random
But they are generated by
a highly random process

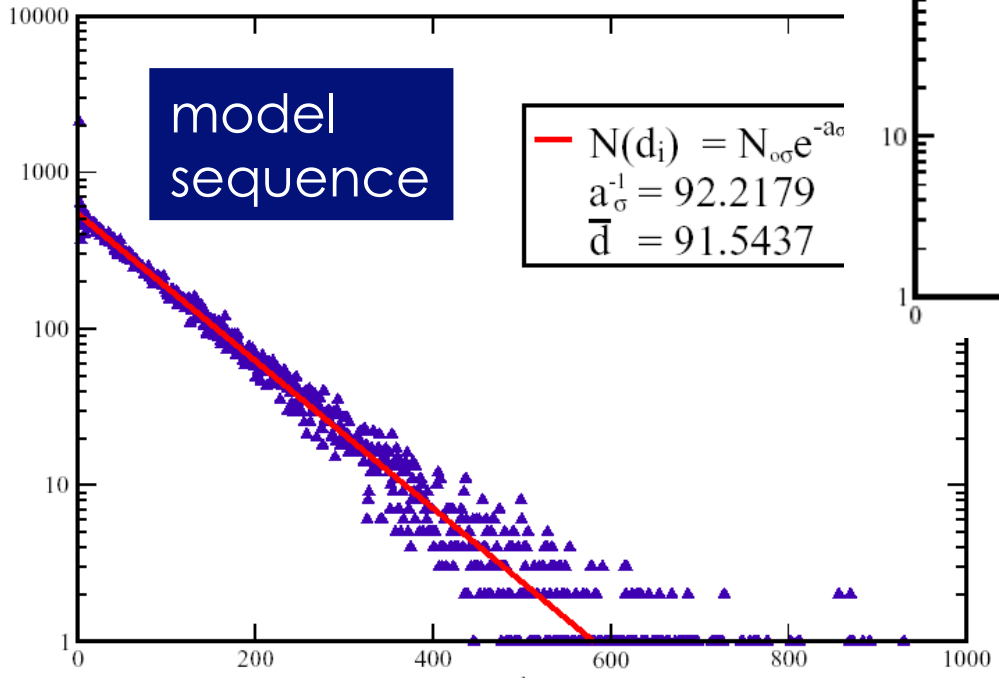
Word Intervals

- Intervals (spatial or temporal) between adjacent random uncorrelated events have an **exponential distribution**
- In a random sequence, intervals of identical words are exponential
- What is the word-interval distribution in a (non-random) genome?



Interval distribution is exponential in random sequence as expected.
But also in genome!

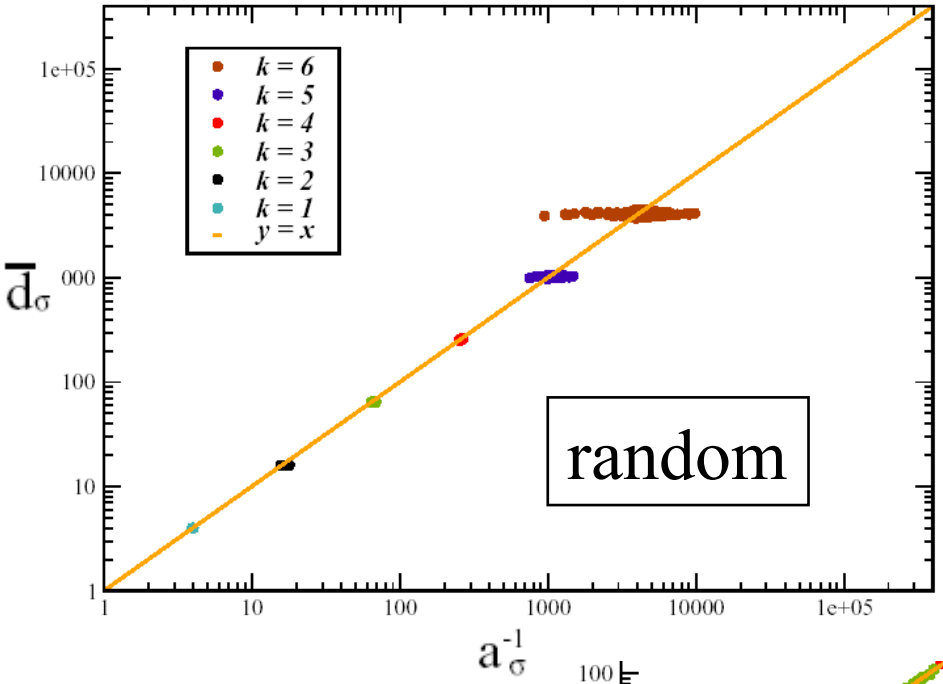


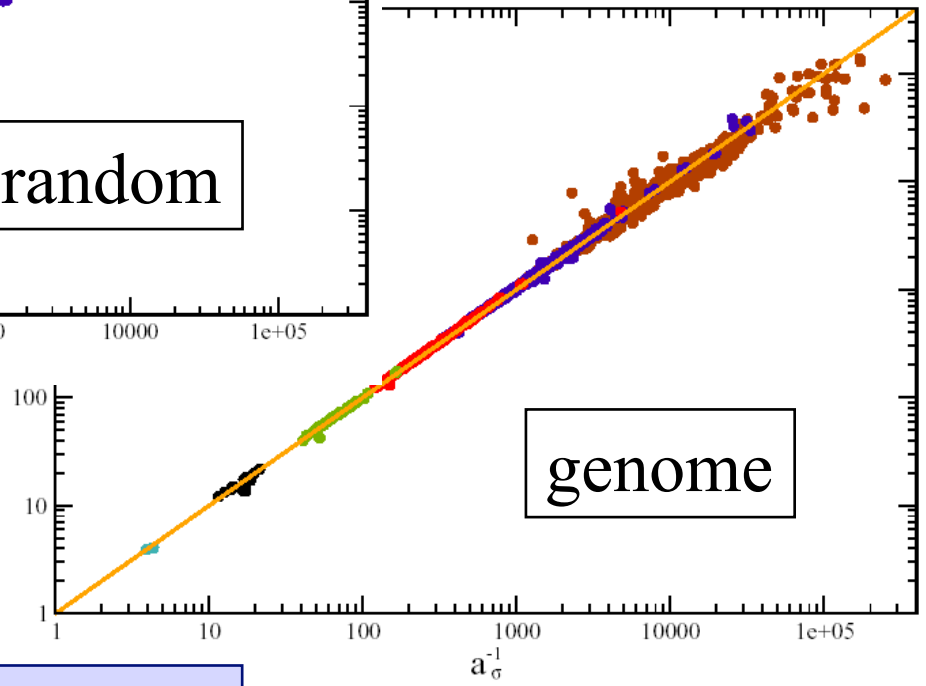


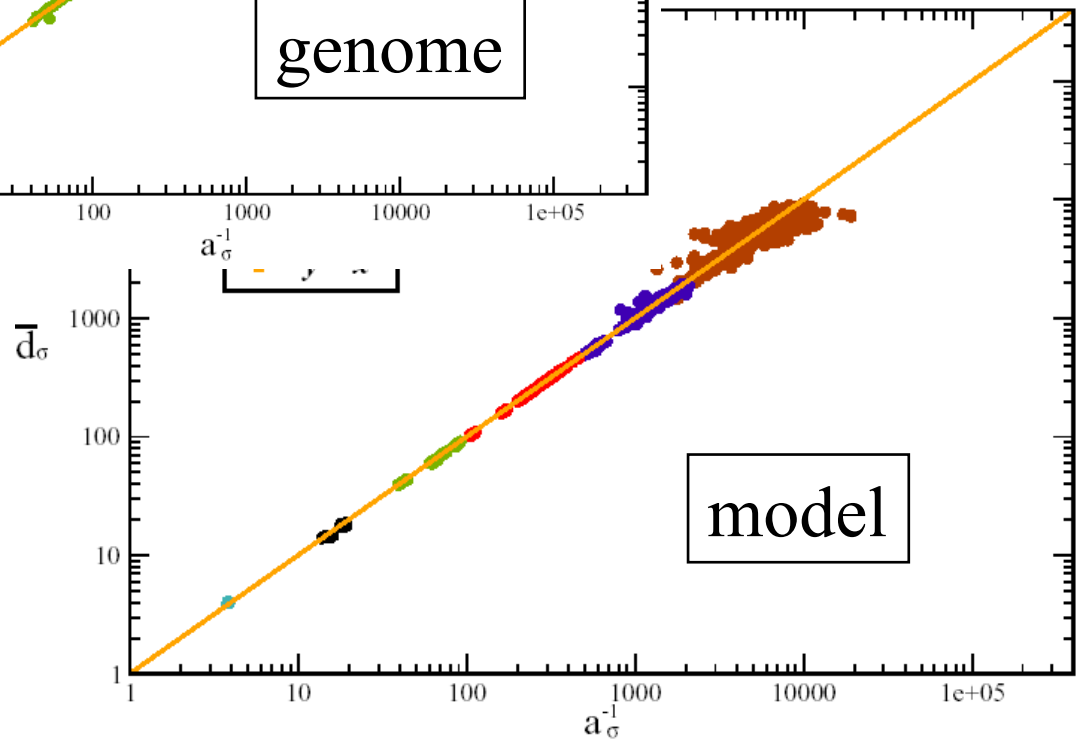
And in the model sequence (not surprising, because growth mechanism is **maximally stochastic**).

All *k*-mers in *E. Coli*, *k*=1-6.

$N(d) = N_0 \exp(-a\,d)$

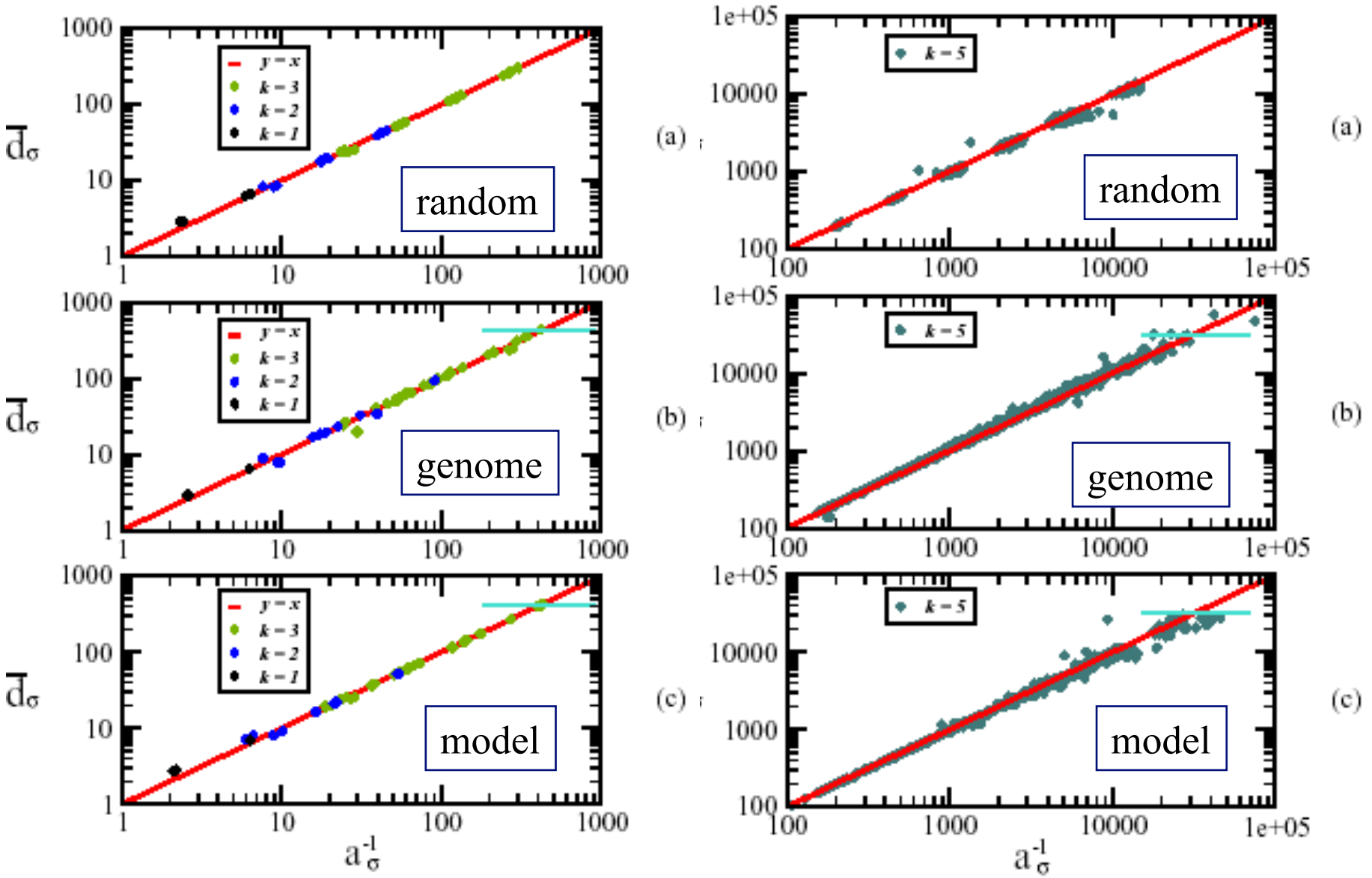






Each data point from one *k*-mer; each *k* has 4^k pts. $\overline{d}_\sigma$ is average d from sequence. In ran. sequence for each *k*, all $\overline{d}_\sigma$ are the same.

For biased composition (p not 0.5), data concentrated at $k+1$ points for each k, but are spread out in genomes.

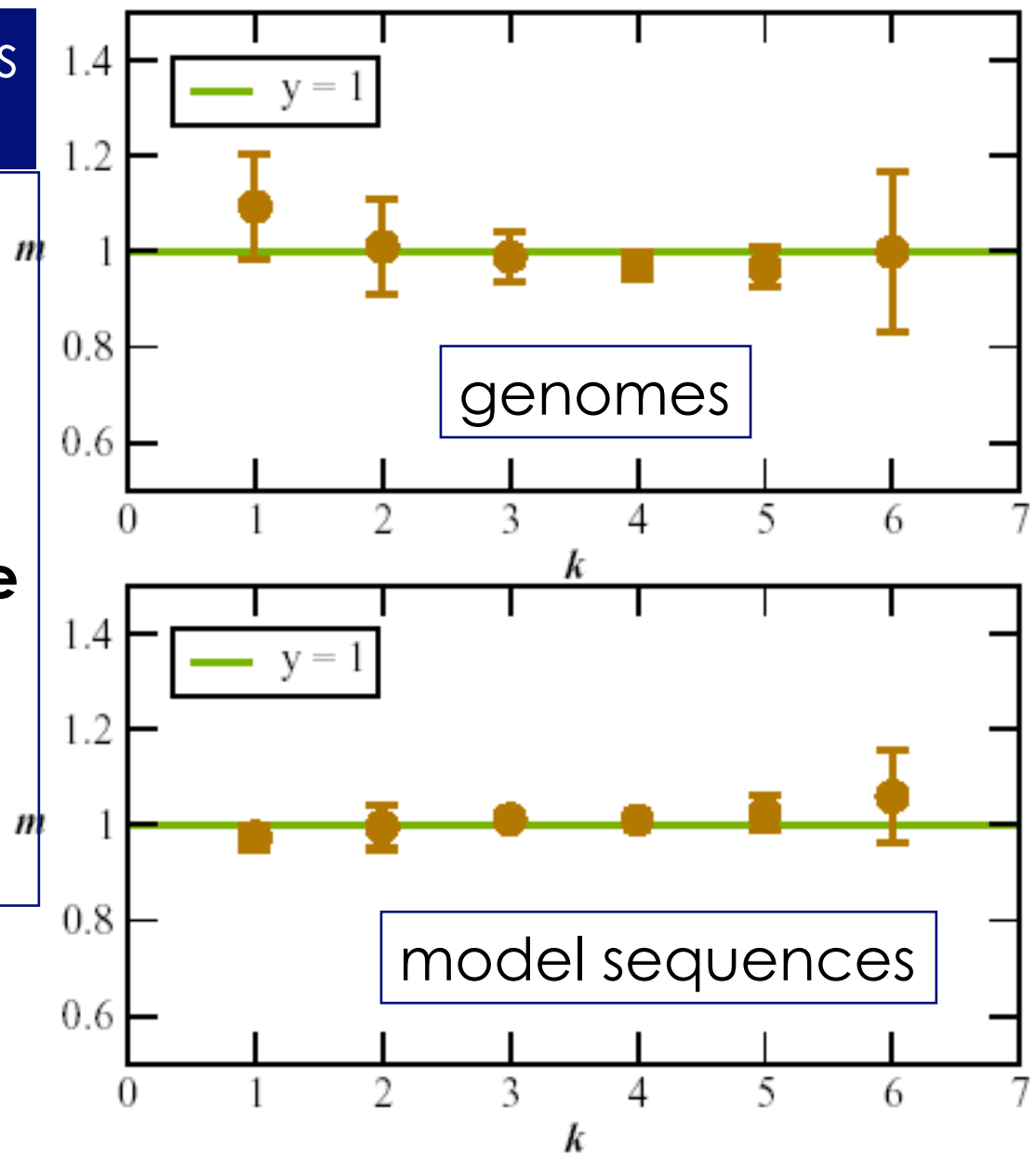


41 microbial genomes longer than 4 Mb

$m = a\,\overline{d}_\sigma$

A from exponential Fit; $\overline{d}_\sigma$ is average *d* from sequence.

Conclusion: words are randomly generated in genomes. Emulated by growth model.



Evolution rates

~ Putting time in our model

Rates & sequence similarity

- Identify substitutions and duplications by sequence similarity ("blasting")
- Substitution rate
  - K: substitution per site between two homologous sequences
  - T: divergence time of two sequences
  - Subst. rate $r_S = K/2T$ (/site/unit time)
- Duplication rate
  - N: number of duplication events per site
  - Duplication rate $r_D = N/T$ (/site/unit time)

Some data on rates for human

- Data
  - Estimated silent site substitute rates for plants and animals range from 1 to 16 (/site/By) (Li97)
  - Humans: r_S ~2 (Lynch00) or 1 (Liu03) /site/By .
  - Animal gene duplication rate ~ 0.01 (0.002 to 0.02) per gene per My (Lynch00)
  - Human (coding region ~ 3% of genome) translates to 3.9/Mb/My.
  - Human retrotransposition event rate ~ 2.8/Mb/My (Liu03)
- Estimate rates for human

$$r_S \sim 2 \text{ /site/By}, \quad r_D \sim 3.4 \text{/Mb/My}$$

- Human genome grew 15-20% last 50 My (Liu03)
- References
  - Lynch & Conery Science 290 (2000)
  - Liu (& Eichler) *et al*. Genome Res. 13 (2003)

Rates from growth model

- Arguments
  - Can estimate substitution and duplication rate if assign total growth time
  - Human genome still growing last 50 My
  - Hence assume total growth time for human genome $T \sim 4$ By
- Get rates average over T

  $\langle r_S \rangle \sim 0.25/site/By, \quad \langle r_D \rangle \sim 0.50/Mb/My$

- About 7~8 time smaller than **recent** sequence divergence estimates

Bridging the two estimates

- Rates are **per length**; hence lower when genome is shorter
- Sequence divergence rates $r_{S,D}$ for last DT~50 My are terminal rates
- Model rates $\langle r_{S,D} \rangle$ averaged over whole growth history, **hence $\langle r_{S,D} \rangle$ less than $r_{S,D}$**
- Assume constant (intrinsic) rate r_C and genome grew exponentially with time

$$L(t) = L_0 \, exp(T/\tau)$$

Bridging … (continued)

- Number of events in interval dt at time t is

$$dN(t) = r_0 L(t)\, dt$$

- $\langle r \rangle$ is average over whole T, r is average over last $\Delta t \sim 0$
- Have $\tau/T << 1$ (because $\langle r \rangle / r << 1$) and $\Delta t/\tau << 1$,
- Then

$$\boldsymbol{r \sim r_0, \quad \langle r \rangle \sim r_0\, \tau/T}$$

- Then from $\tau/T \sim \langle r \rangle / r \sim 1/8$

$$\boldsymbol{\tau \sim 0.5\ By, \quad L_0 \sim 1\ Mb.}$$

Human rates and growth (summary)

- Very roughly, constant rates in human
  - site substitution: $r_S \sim 2$ /site/By,
  - segmental duplication $r_D \sim 3.4$/Mb/My,
- Growth
  - $L(t) \sim 0.001$ (Bb) $L_0 \exp(t/0.5$ (By))
- Remarks
  - grew by ~ 12% last 50My
  - Liu *et al.* grew by ~ 15-19% last 50My
  - Does not imply L=1 Mb at t=0
  - Does imply at $t << 500$My, L ~ 1 Mb

Discussion & Implications

~ *Genomes are close to being self-organized critical systems*

~ *Evolution mostly driven by neutral events*

Summary of results

- Genomes are large systems with small-system statistics
- Shannon information of complete genomes exhibit **universal lengths;** genomes belongs to single **universality class**
- Data consistent with simple growth model based on **maximally stochastic segmental duplication and random point mutation**
  - For **human genome**, site substitution and segmental duplication **per site per time rates** consistent w/ those extracted by sequence divergence methods
- Genomes are not random but are essentially **randomly generated**
  - Has high degree of self-similarity, almost **SOC** systems
- Model permits **universal or multiple ancestor** as well as **huge species diversity**

Neutral theory of evolution

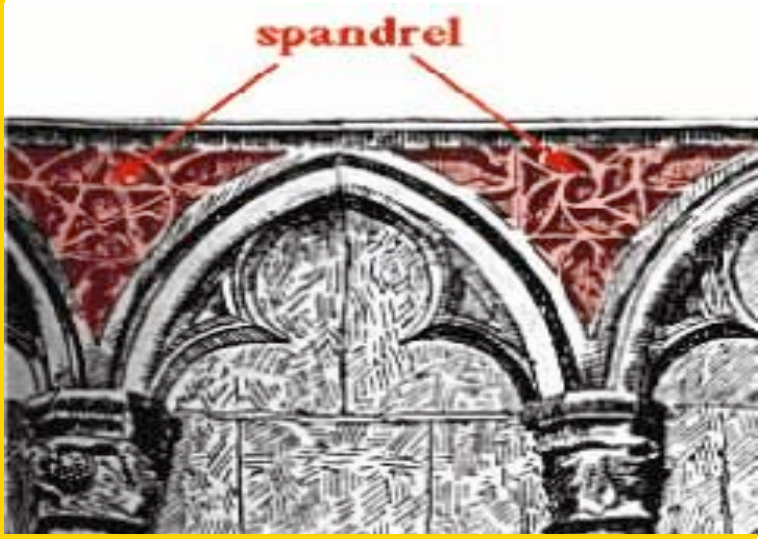
- Stochastic Duplication/replication was superb evolutionary strategy
  - A most efficient way to:
  - Grow and accumulate information
  - Escape rule of large systems
- Duplication/replication and mutations were mostly selectively neutral
  - because measure not sensitive to coding
  - most of eukaryotic genomes are non-coding parts
  - Eukaryotes and prokaryotes belong to the same universality
- Corroborates Kimura's neutral theory of molecular evolution (1968, 1983)
  - based on polymorphisms of genes
  - most mutations on genes were selectively neutral

Shannon information *versus* biological information

- Large Shannon information is necessary condition for rich biological information
- Growth by random duplication provides an basis allowing natural selection to fine-tune, via natural selection, Shannon information into biological information
- The adaptation of the strategy of growth by random duplication by itself may be a consequence of natural selection

Are genes "spandrels"?

- Spandrels
  - In **architecture**. The roughly triangular space between an arch, a wall and the ceiling
  - In **evolution**. Major category of important evolutionary features that were originally side effects and did not arise as adaptations *(Gould and Lewontin 1979)*



- The duplications may be what the arches, walls and ceilings are to spandrels and the genes are the decorations in the spandrels

Classical Darwinian Gradualism or Punctuated equilibrium?

- Great debated in palaeontology and evolution - Dawkins & others vs. (the late) Gould & Eldridge: evolution went gradually and evenly vs. by stochastic bursts with intervals of stasis

Our model provides genetic basis for both. Mutation and small duplication induce gradual change; occasional large duplication can induce abrupt and seemingly discontinuous change

The RNA World

- RNA was discovered in early 80's to have enzymatic activity – ribozymes can splice and replicate DNA sequences *(Cech et al. (1981), Guerrier-Takada et al. 1983)*
- The RNA world conjecture – early had no proteins, only RNAs, which played the dual roles of genotype and phenotype
- Some present-day ribozymes are very small; smallest hammerhead ribozyme only 31 nucleotides; ribozymes in early life need not be much larger

RNA World & size of early genome

- In our model the small initial size of the genome necessarily implies an early RNA world
- A genome 200~300 nt long is long enough to code the many small ribozymes (but not proteins) needed to propagate life
- Origin of this initial genome not addressed in the model. It (or its presursor) could have arisen spontaneously - artificial ribozymes have been succcessfully isolated from pools of random RNA sequences *(Ekland et al. 1995)*
- Present-day ribozyme can be as small as 31 nt; there could be smaller earlier ones.

Growth by duplication may provide partial answers to:

- How did life evolve so rapidly?
- How have genes been duplicated at the high rate of about 1% per gene per million years? *(Lynch 2000)*
- Why are there so many duplicate genes in all life forms? *(Maynard 1998, Otto & Yong 2001)*
- The chromosome exchanges that characterize mammalian and plant radiations. *(O'Brien et al. 1999; Grant, et al. 2000)*
- Was duplicate genes selected because they contribute to genetic robustness? *(Gu et al. 2003)*
  - Likely not; Most likely high frequency of occurrence duplicate genes is a spandrel

Many more questions to answer

- Tracing natural selection
  - Can we show conclusively growth by stochastic duplication is faster than selection driven (at gene level) growth?
  - Can we extend the method to say anything about evolution of genes? (Introduce roughness in genome?)
- Time scale
  - When did growth happen? At what rate? How did growth stabilize? Has it stabilized?
  - When and how did the codons form? When did protein arise?
- Phylogeny
  - Is the model useful for using whole genomes to build trees?
  - If so will the result agree with alignment bases analysis?
- Universal ancestor
  - Was there a Universal Ancestor ? Or were there a group of Ancestors?

Collaborators & Acknowledgement

- Collaborators
  - Professor Liaofu Luo (Univ. Inner Mongolia)
  - From Comptational Biology Lab (Nat'l Central U.)
    - 謝立青博士 Dr. LC Hsieh
    - 陳大元（博生）TY Chen (PhD Candidate)
    - 陳鴻大（博生）HD Chen (PhD Candidate)
    - 張昌衡（碩生）CH Chang (MSc Candidate)
    - 范文郎 （碩生）WL Fan (MSc Candidate)
- Thanks to:
  - National Center for Theoretical Science (Shinchu)
  - Institute for Theoretical Physics (CAS,Beijing)
  - Center for Theoretical Biology (Beijing U.)

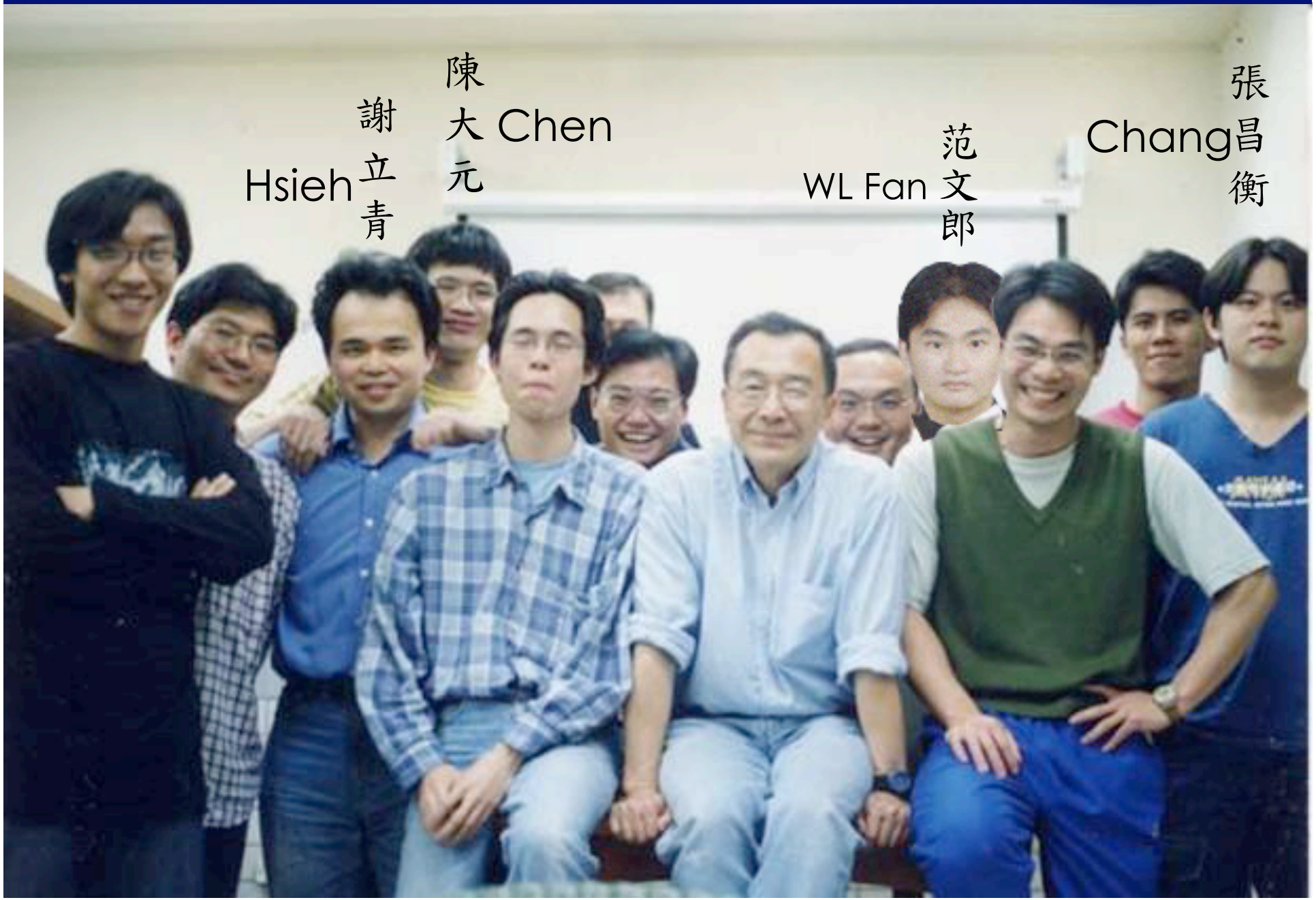
Some references

- Model for growth of bacterial genomes, LS Hsieh and HCL, Mod. Phys. Lett. 16 (2002) 821-827
- Short Segmental Duplication: Parsimony in Growth of Microbial Genomes, LS Hsieh, LF Luo and HCL, Genome Biology, 4 (2003) 7
- Minimal model for genome evolution and growth, LC Hsieh *et al.*, Phys. Rev. Letts. 90 (2003) 018101-104
- Universality in large-scale structure of complete genomes, LS Hsieh *et al.*, Genome Biology, 5 (2004) 7
- Universal signature in whole genomes, TY Chen *et al.*, (submitted to Science) http://sansan.phy.ncu.edu.tw/~hclee/ppr/hsieh_text.pdf
- Shannon information in complete genomes, CH Chang, *et al.*, (to appear in Proc. IEEE Computer Society Bioinformatics Conference (CSB2004))

  http://sansan.phy.ncu.edu.tw/~hclee/rpr/Lee_H_Shannon.pdf

For copies, see http://sansan.phy.ncu.edu.tw/~hclee/pub/selected.html

Computation Biology Laboratory (2003)



謝謝

Thank you for your attention

RNA World & length of duplicated segments

- Present-day ribozyme can be as small as 31 nt; there could be smaller earlier ones.
- The average duplicated segment length of 25 nt in the model is very short compared to present-day genes that code for proteins, but likely represents a good portion of the length of a typical ribozyme encoded in the early universal genome of the RNA world