Complexity, Universality, Self-similarity & Growth of Genomes

Santa Fe Institute
CSSS 2005

Xiangshan, Beijing, China
2005 July 10 - August 5

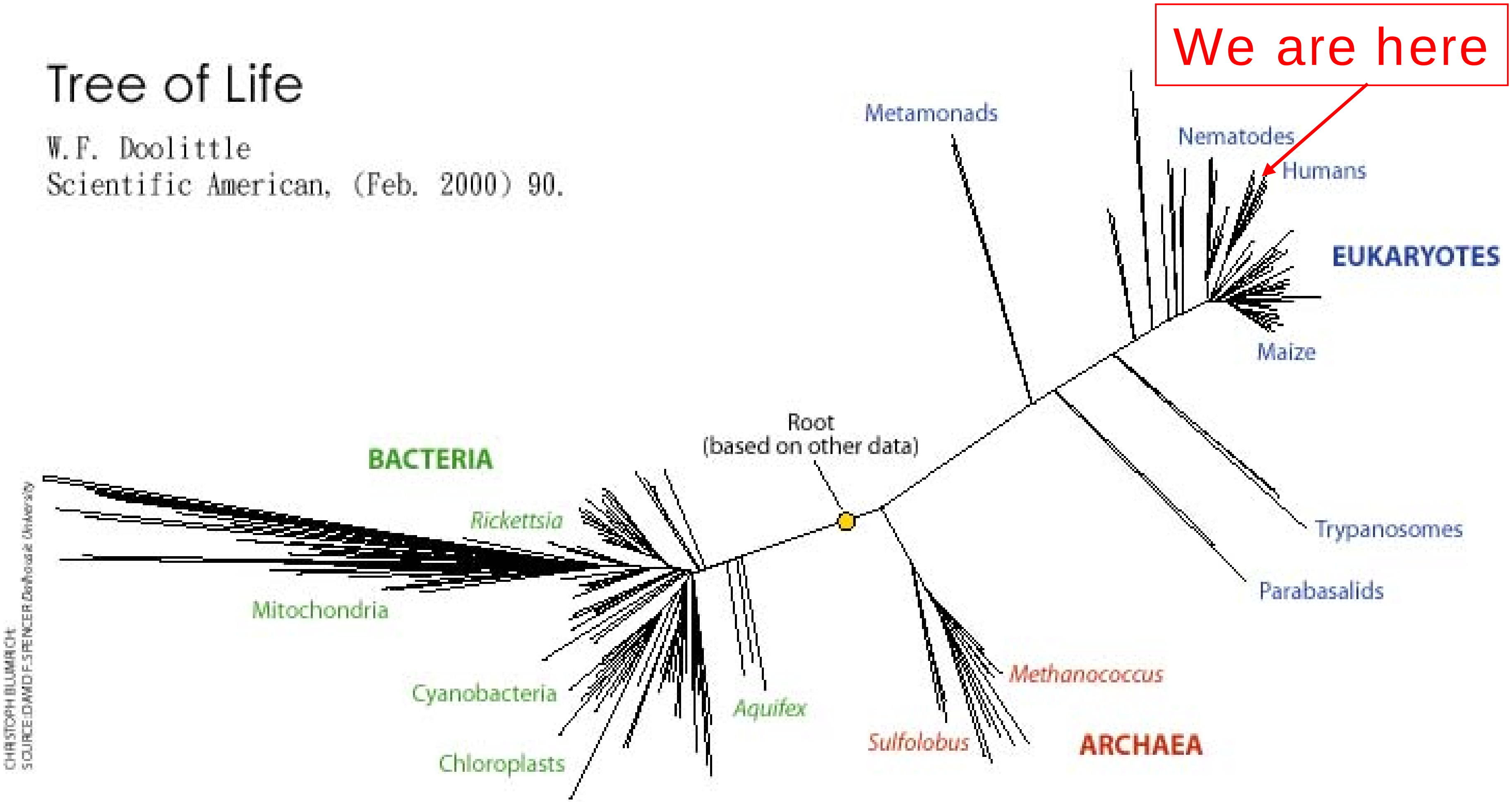
HC Lee

Computational Biology Lab
Dept. Physics & Dept. Life Sciences
National Central University

Some concepts to be discussed

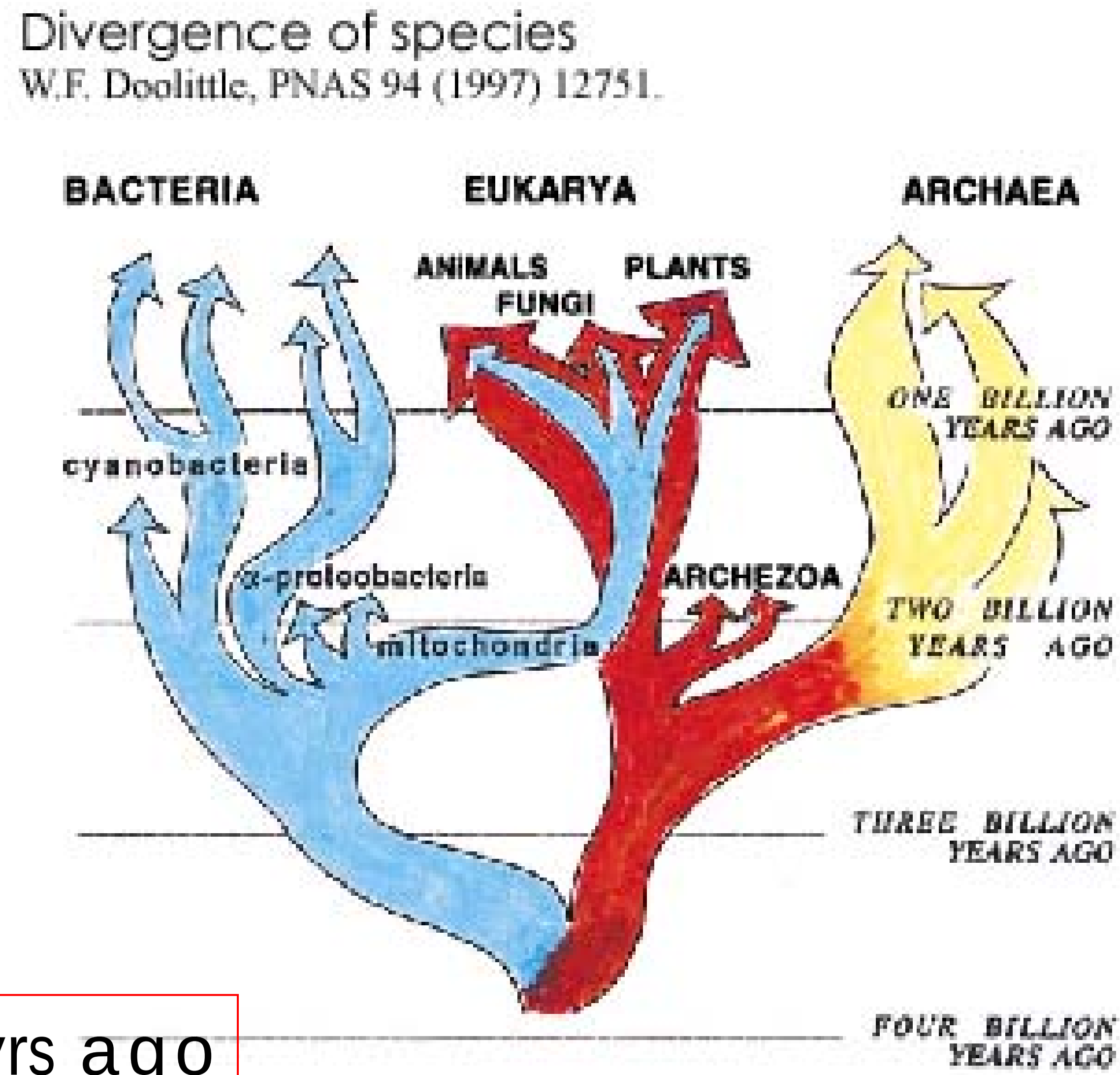
- By examining at the textual property of genomes, we encounter/exploit the following concepts
  - Randomness and order
  - Second law of thermodynamics
  - (Shannon) Information and entropy
  - Distribution and its variance
  - Diversity and universality
  - Complexity and self-similarity
  - Neutral evolution and natural selection
- and arrive at a hypothesis and model for genome growth

Life is highly diverse and complex



And it took a long time to get here

now



4 billion yrs ago

Evolution of life is recorded in genomes

- Genome is Book of Life
- A double helix - two strands of DNA
- DNA: String of four types of
- molecules – chemical letters - A, C, G, T
- Genome is a linear text written in four letters
- We believe all genomes have a common ancestor, or a small group of ancestors

Genomes are BIG

A stretch of genome from the X chromosome 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 aggcccctct 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 tattttttct 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
```

Evolution of Genomes and the Second Law of Thermodynamics

Genomes grew & evolved 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

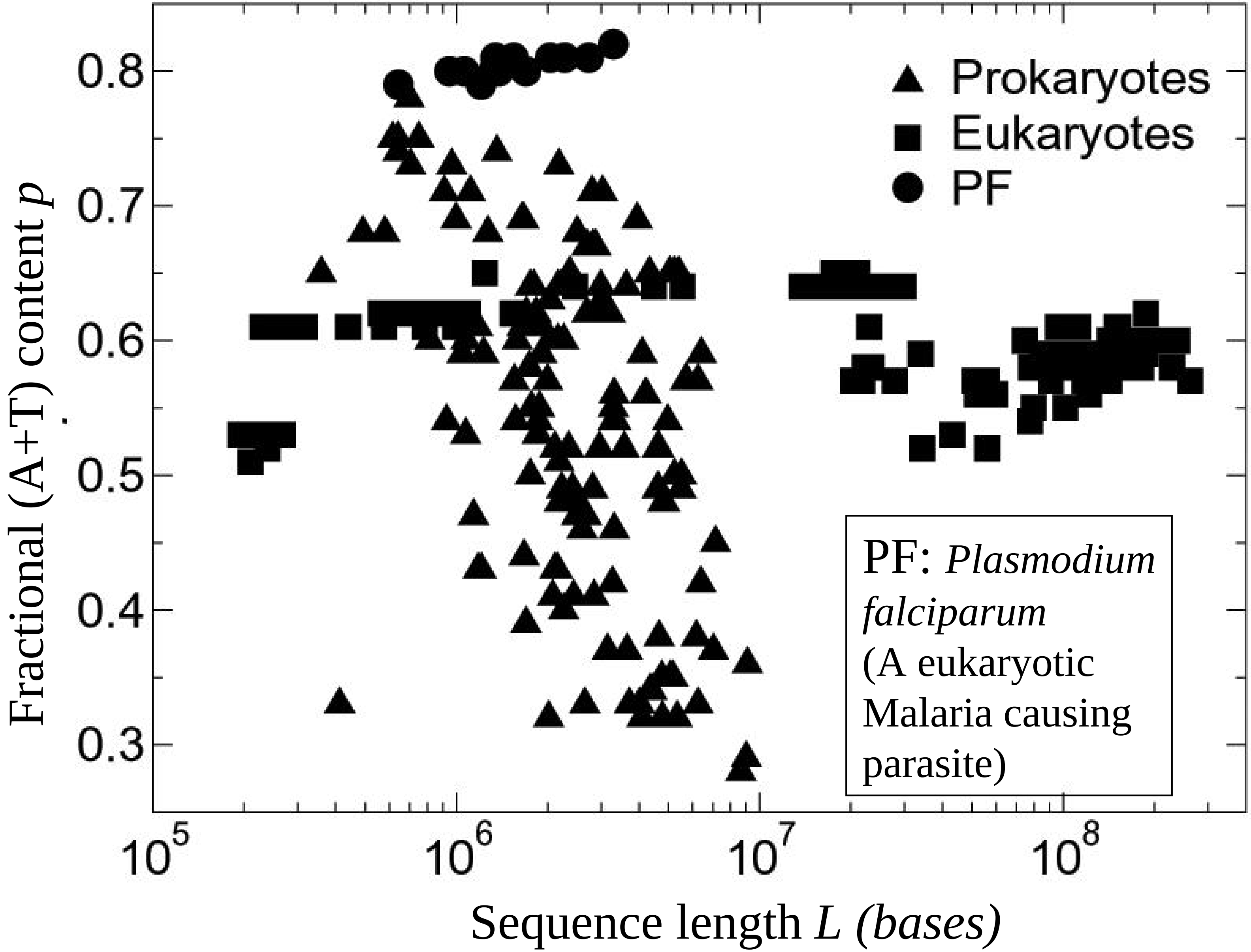
- Shannon information
  - Information decreases with increasing entropy

- How was genome able to simultaneously grow stochastically AND acquire information?

Characterization of Genomes

- Primary characterization of genomes
  - length in bp (base pair)
  - base composition p = A+T/(A+T+C+G)
  - word frequencies
- Secondary characterization
  - % coding region (microbials: ~85%; eukaryotes (2~50%)
  - number of genes (few hundred to 25K)
- Tertiary characterization
  - intron/exon (microbials, no; eukaryotes, yes)
  - other details

Complete Genomes are diverse



Distribution and Width

- Consider τ equally probable events occurring a total of L times.

- Distribution of occurrence frequency characterized by
  - mean frequency: $f_{ave} = L/\tau$
  - SD (standard deviation) Δ; or
    CV (coefficient of variation) = Δ/f_{ave}
  - Higher moments of distribution

Random events

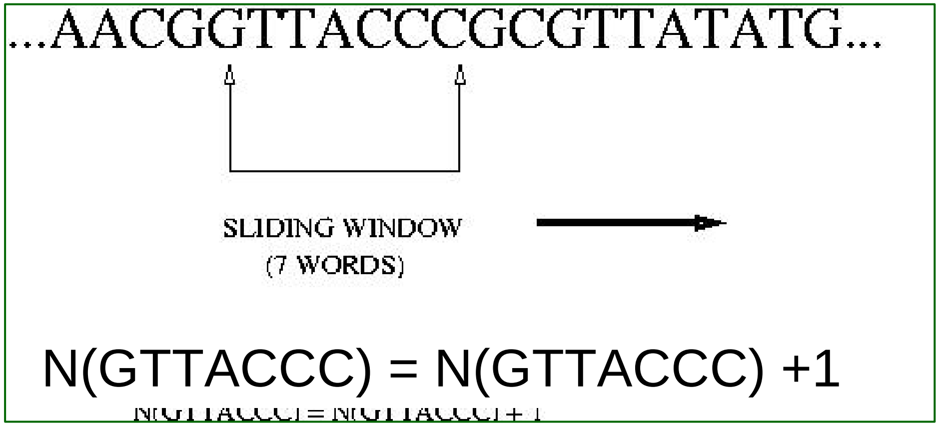
- Random events given by Poisson distribution
  - $\Delta^2 = f_{ave}$, or, $(CV)^2 = 1/ f_{ave}$
  - That is, $(CV)^2 = \tau/L$

- For fixed τ, $(CV)^2 \sim 1/L$
  - Large L limit (thermodynamic limit): $L \sim$ infinity, $CV \sim 0$

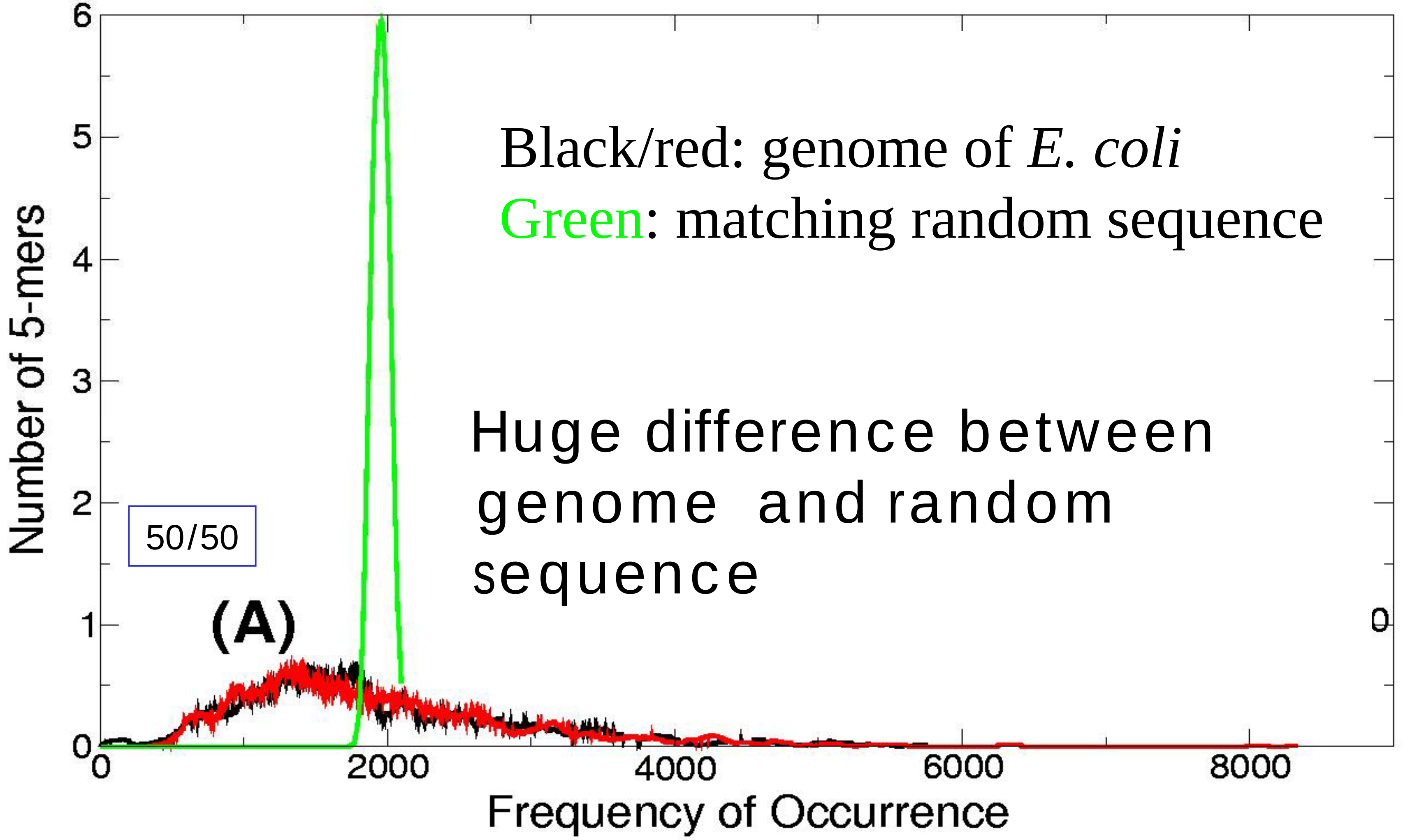
- For given τ, if CV is known, then
  - $L \sim \tau/(CV)^2$

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 $\{f_i \mid i=1 \text{ to } 4^k\}$



For genomes: events=word occurrence; type of events τ=types of words = 4^k; distr.= distr. of frequency of occurrence



Two big surprises from complete genomes

Given τ and CV, define effective length

$$L_{eff} = \tau/(CV)^2$$

- The L_{eff} of complete genomes are far shorter than their actual lengths
- For a given type of event (word counts) L_{eff} is universal
  - Actual length varies by factor > 1000
  - "Information" in genomes growths as L

Large CV, or small L_{eff}, implies more "information"

Compare L_{eff} with true length L for all complete genomes for 2-10 letter words

$$(CV_{genome})^2 = \tau / L_{eff}$$

def

$$(CV_{random})^2 = \tau / L$$

$$M_s = (CV_{genome})^2 / (CV_{random})^2 = L / L_{eff}$$

Note: technical details when *p* not equal to 0.5

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 & coefficient of variation

- **Shannon information**: information *decreases* with increasing H. Define:

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

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

- Relation to **coefficient of variation** (for unimodal distribution)

$$R \equiv \ln\tau - H(\mathcal{F}) = L^{-1}\sum_i f_i \ln(f_i/\bar{f}) = (CV)^2/2 + \ldots$$

- Shannon information and coefficient of variation are equivalent measures

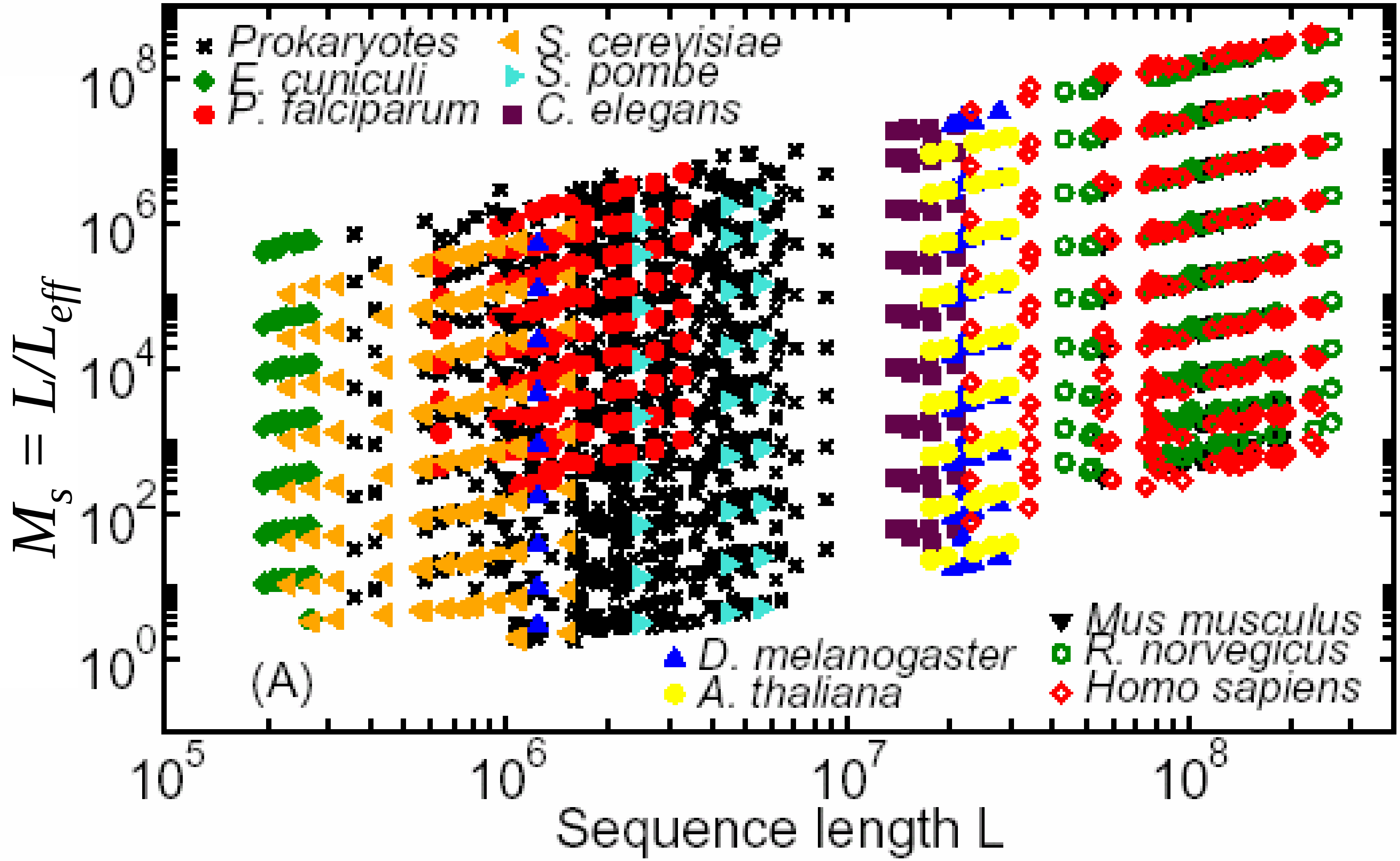
(Note: technical detail re biased base composition important)

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

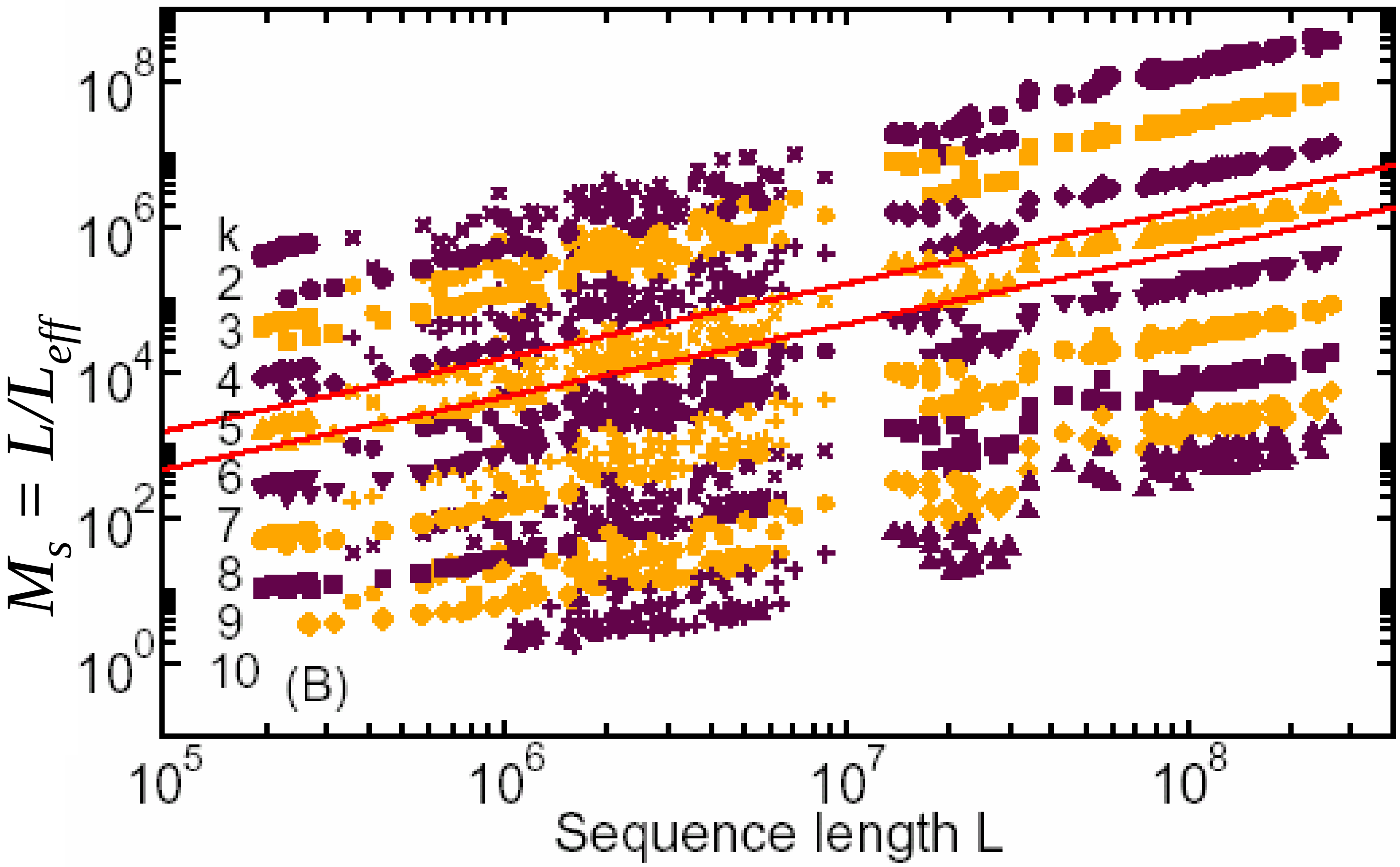
	Random sequence			Genome sequence		
k	H	R	R_{ex}	H	R	R_{gen}/R_{ran}
2	3.9999	5.90 E-6	5.77 E-6	3.973	2.66 E-2	**4500**
3	5.9999	3.72 E-5	3.46 E-5	5.933	6.65 E-2	**1922**
4	7.9999	1.72 E-4	1.62 E-4	7.881	1.18 E-1	**728**
5	9.9993	7.26 E-4	7.53 E-4	9.821	1.79 E-1	**246**
6	11.999	2.94 E-3	2.90 E-3	11.75	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	**-**

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*



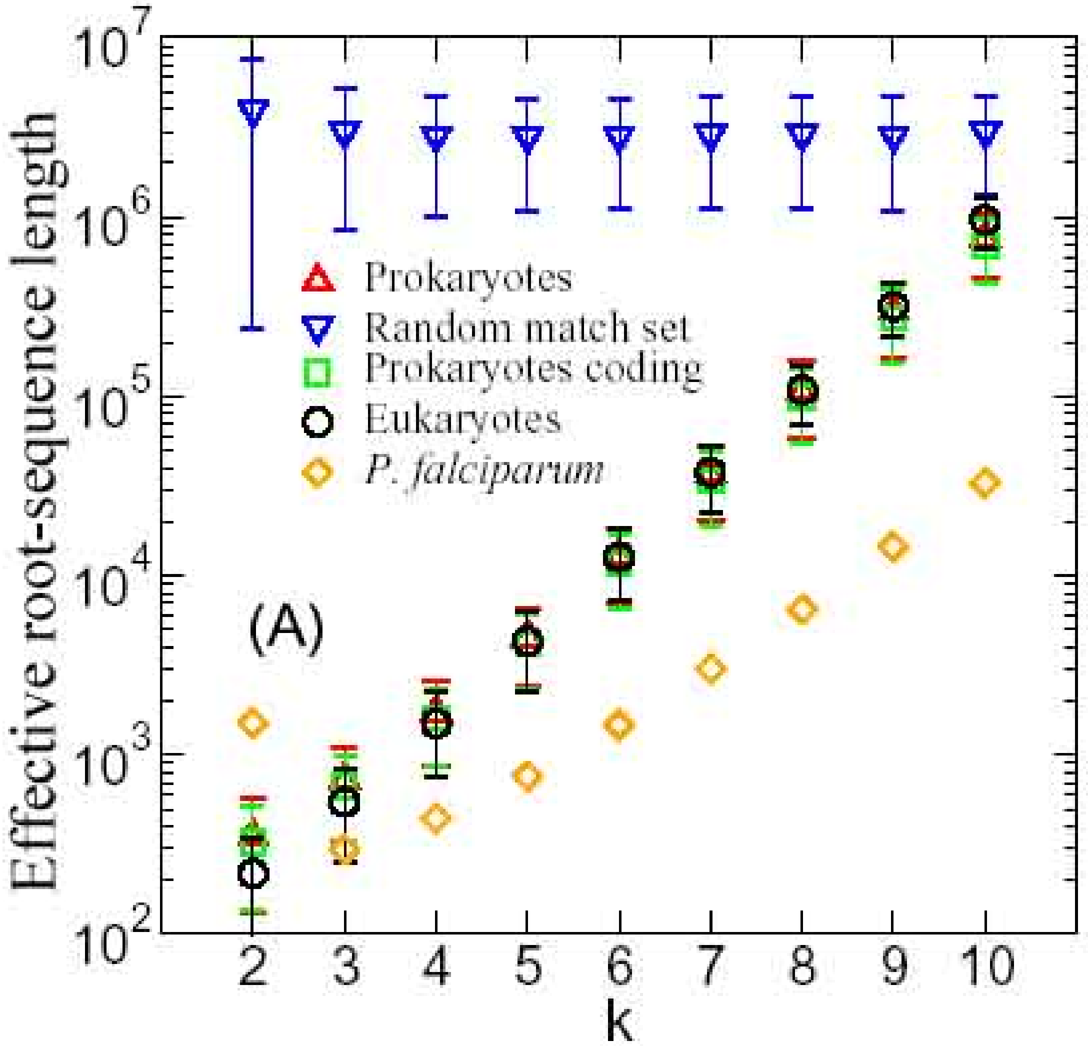
Data from 14 *Plasmodium* chromosomes excluded; ~2400 data points. For each k, 268 data points form a narrow $M_\sigma \sim L$ "k *band*".

Genomes are in Universality Classes

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

  $$\log L_r(k) = a\,k + B$$

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



Black: genome data; green: artificial

Self-similarity

- Self-similarity: “Sameness” at varying scales
- We have seen: complete genome of length L has stats property of random sequence of length $L_{eff} \lll L$
  - *Question 1: what is the stats property of a $L' > L_{eff}$ segment of the genome?*
  - *Question 2: what happens when we concatenate two such segments?*

Testing self-similarity

	Behavior of $(CV)^2$	
	Random sequence	Genome
Any segment $L >> L' >> L_{eff}$	$\sim 1/L'$	$\sim 1/L_{eff}$
Concatenation of two segments $L >> L_1, L_2 >> L_{eff}$	$\sim 1/(L_1 + L_2)$	$\sim 1/L_{eff}$

Two examples: *H. sapien* and *E. coli*: genomes are highly self-similar

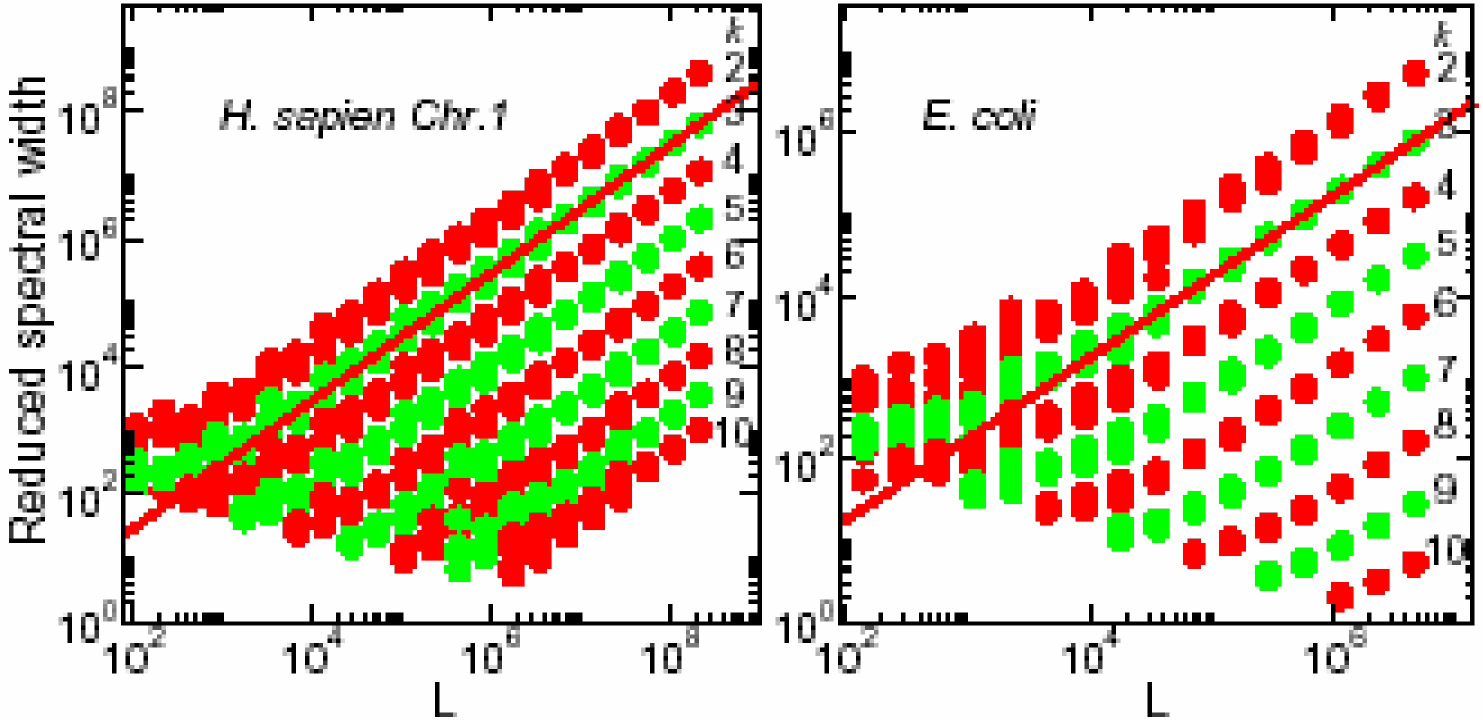


Figure 1: RSW ($\mathcal{M}_\sigma$) of k-spectra, k=2 to 10, of segments from the 246 Mb chromosome I of *H. sapiens*. Lengths of the segments are $1/2^n$ of full length, n=1 to 21, and for each length eight segments are randomly selected. Data for which segment length is less than 4^k are not included. 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.

L_u and L_d, k=5, all complete sequences

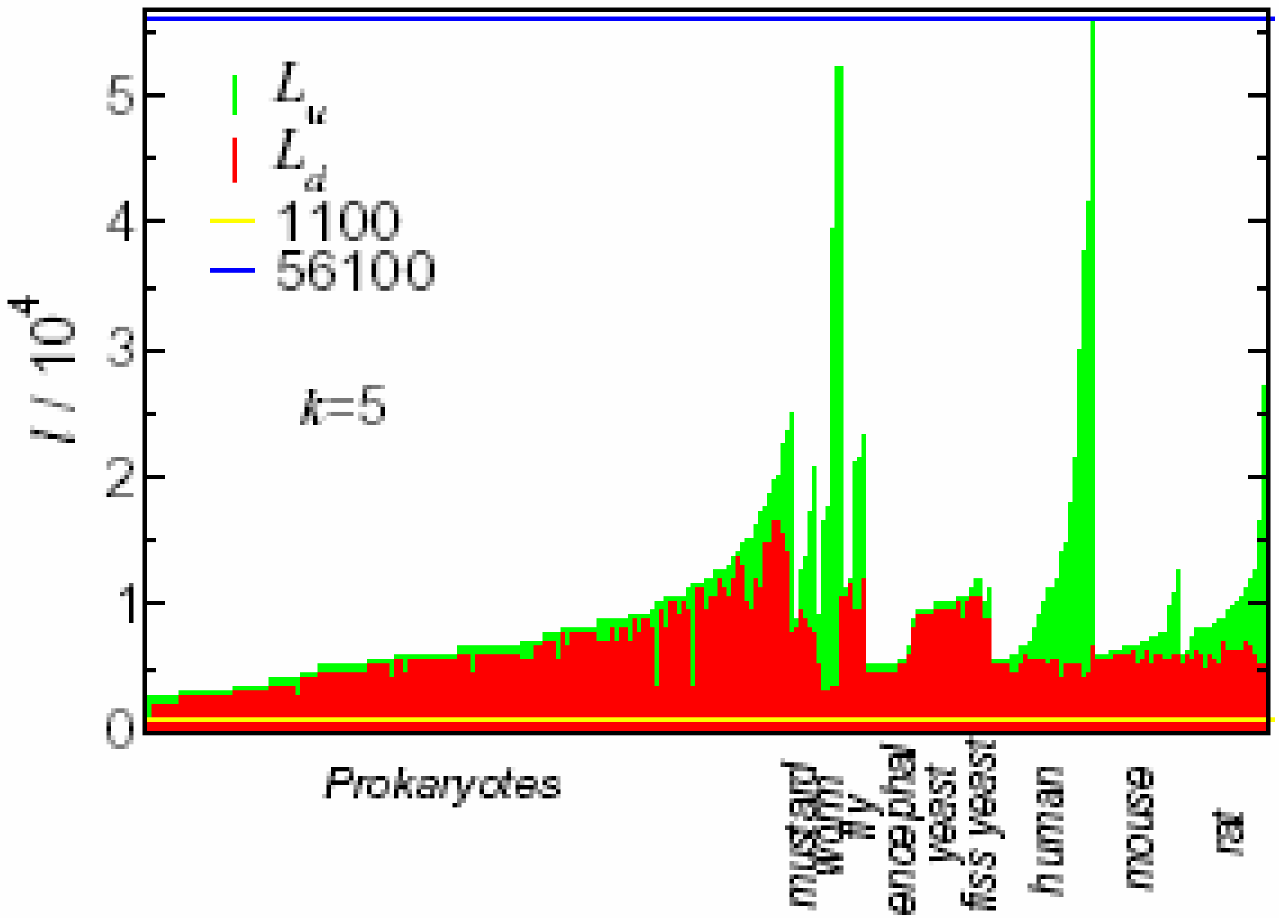


Figure 3: L_u (the length above which all segments are similar to the genome; green bars) and L_d (the length below which no segment is similar to the genome; red bars) for k=5 for all complete sequences in the main universality class. The blue (yellow) line is the position of L_{max} (L_{min}).

Genomes are maximally self-similar

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 & L_d & eukaryotic L_d
- L_{sim} barely > L_r barely > 4^k,
- Hence **genomes are almost maximally self-similar**

Event intervals

- Given a sequence of events. Consider the distribution of intervals between adjacent events
  - Random events: distribution is exponential *[d_{ave}= average interval]*

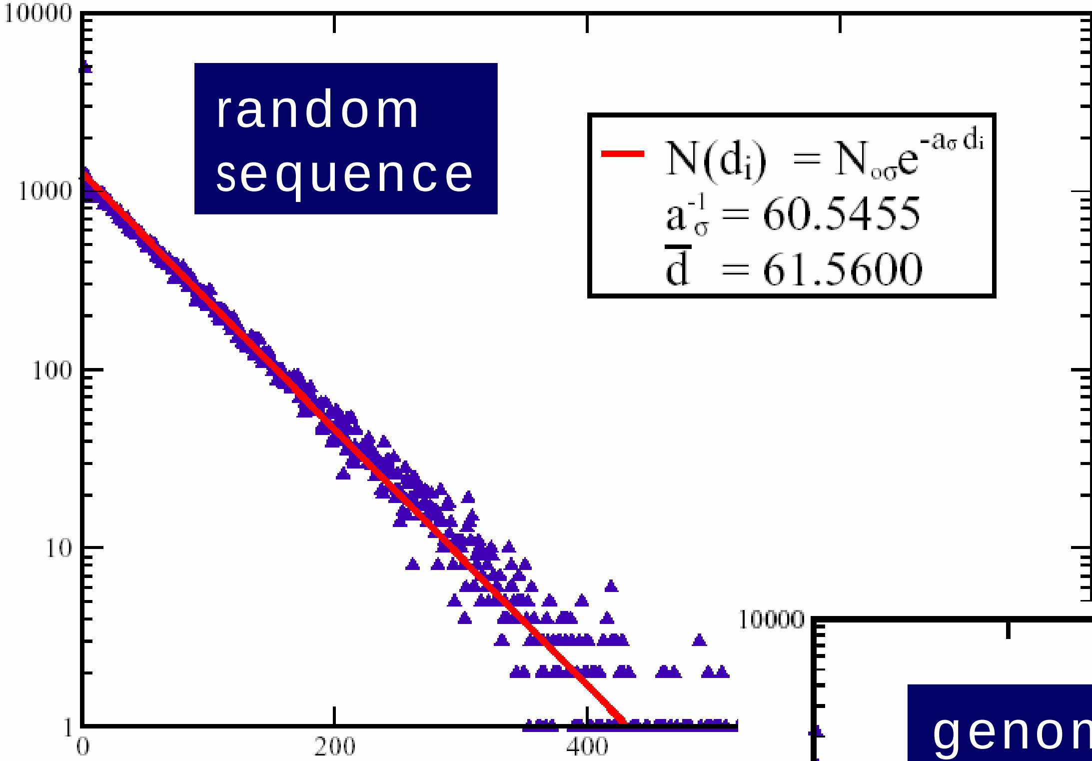
    $$N(d) \sim N_0 \exp(-d/d_{ave})$$

  - Conversely, if distribution is exponential, then infer events occurred randomly (or *vise versa*)

Words occurred in genome randomly

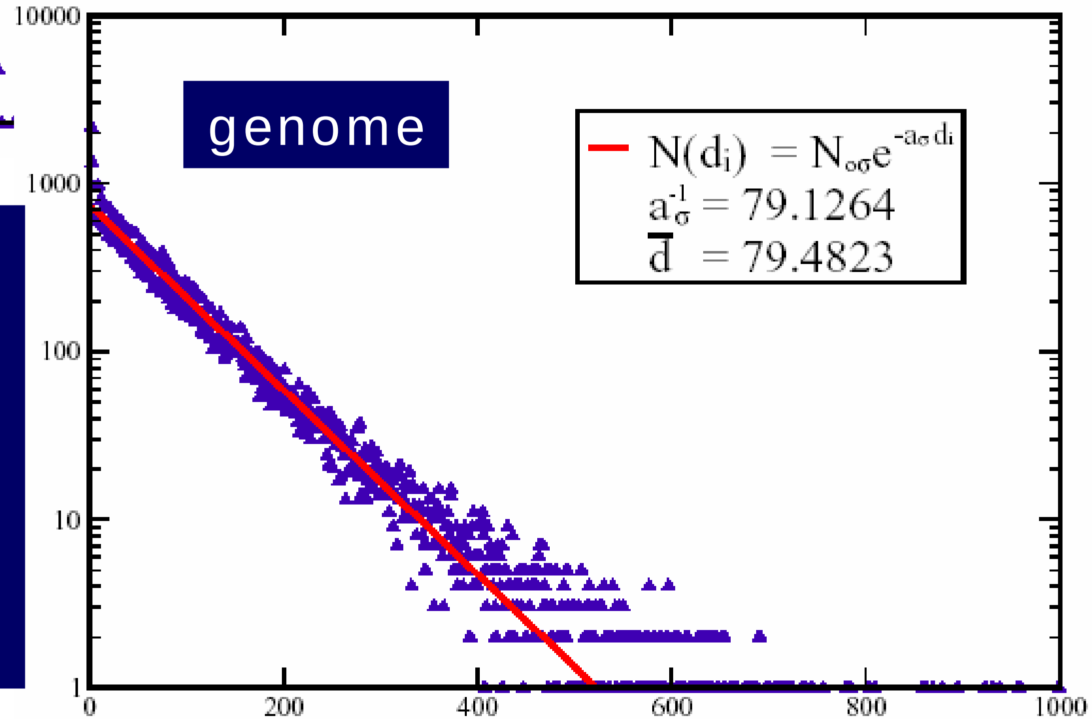
- Have already seen genomes are highly non-random
- Yet, distributions of words intervals in genomes are universally exponential to a high degree of accuracy

Typical interval distribution for a *k*-mer



Interval distribution is exponential in random sequence as expected.

But not so in genome!



Summary of genome data

- Universality class – for fixed word length k, L_{eff} is (approximately) the same for all genomes
  - Log $L_{eff}(k) = ak + B$
    a, B are universal constants
- Maximally self-similar
- k-mer intervals have exponential distribution
- What is the cause of these properties?

Order, Randomness, L_{eff} and duplications

- If we take random sequence of length L_0 and replicate it *n* time, then total sequence length (*L*) is nL_0 but L_{eff} of sequence remains L_0
- Smaller L_{eff} implies higher degree of ORDER
- Larger L_{eff} implies higher degree of RANDOMNESS
- Small L_{eff} of genomes suggests many DUPLICATIONS

A Universal Model for Genome Growth

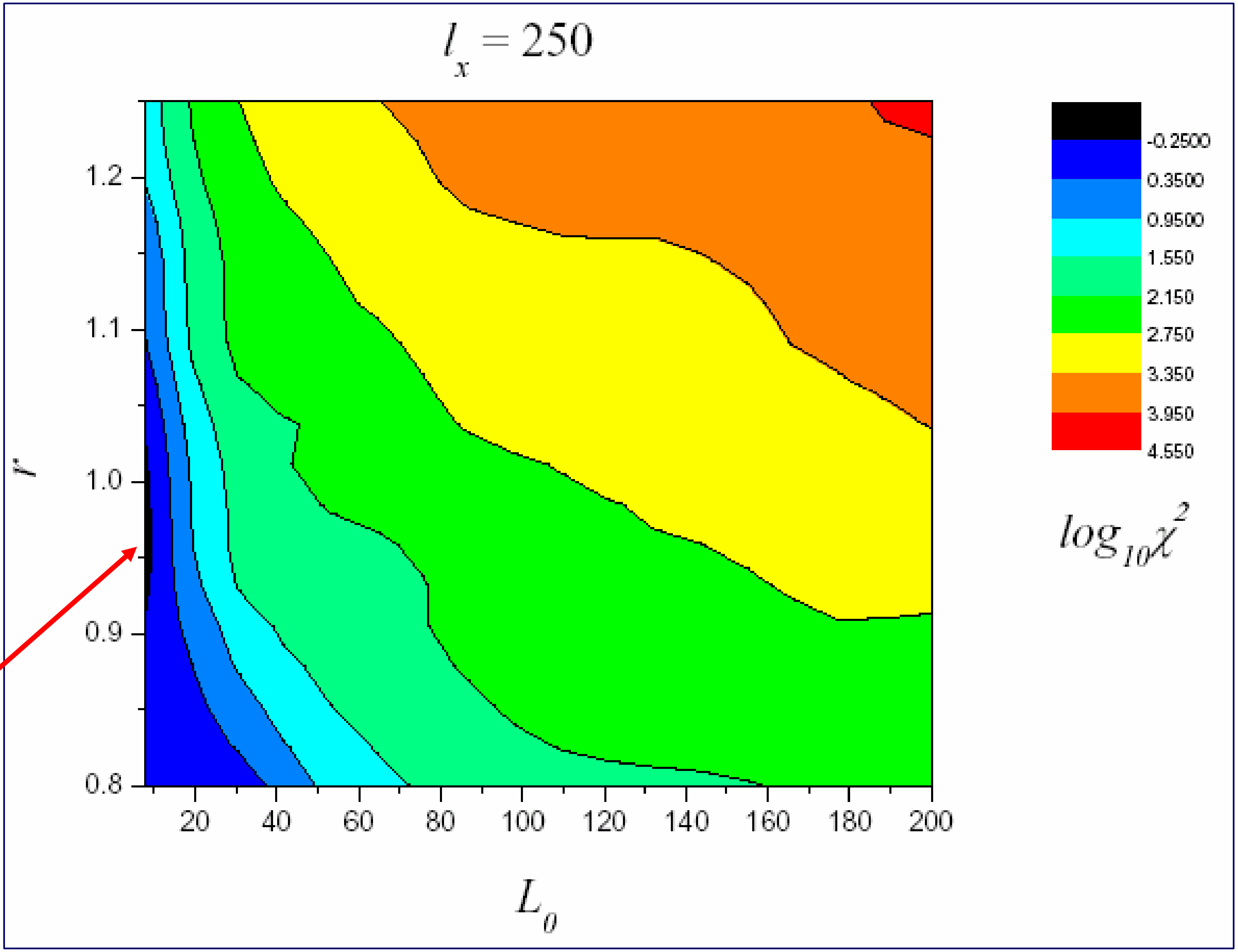
A model: at a universal initial length, genomes grew (and diverged) by maximally stochastic segmental duplication

1. Universal initial length - Common ancestor(?), universal L_{eff}.
2. Segmental duplication – L-independent CV
3. Maximum stochasticity – self-similarity, random word interval

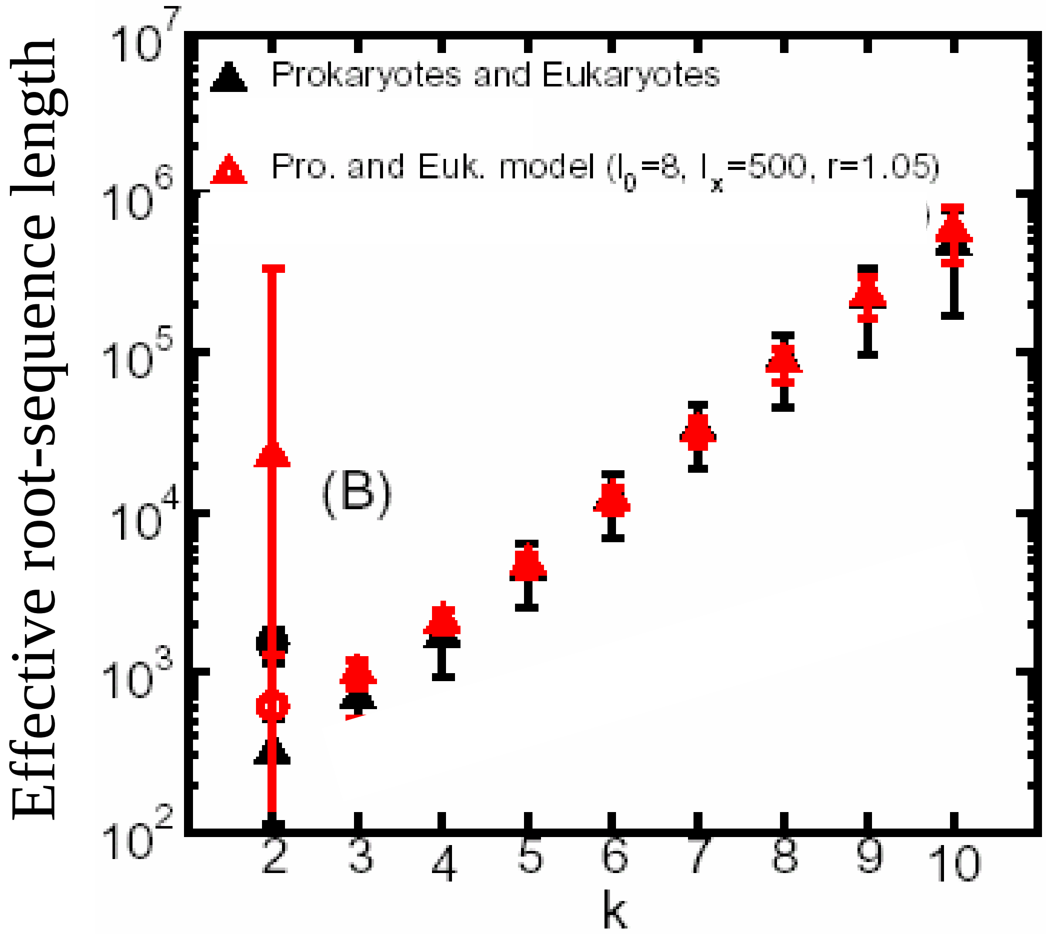
Self copying – strategy for retaining and multiple usage of hard-to-come-by coded sequences (i.e. genes)

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

Model parameter search: favors very small L_0



Model has three universal parameters - generates universal L_{eff}



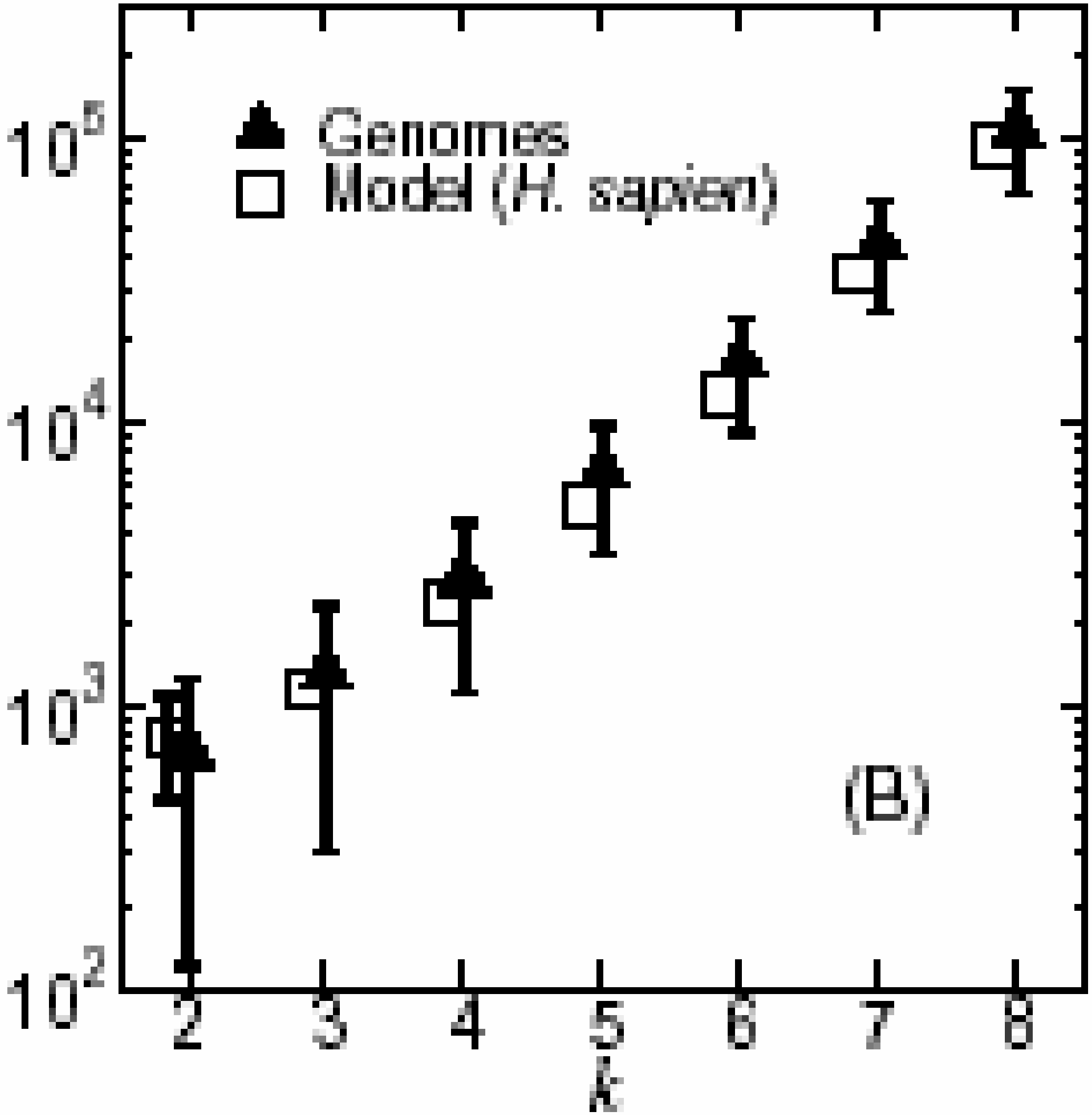
Red & blue symbols are from (same) model sequences

Model sequences are maxiamally self-similar - L_{sim} agrees with data

Note: Model predates data

But model has **smaller spread**

Model is too smooth



Mutation & duplication rates from sequence alignment studies of human genome

- 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
- References
  - Lynch & Conery Science 290 (2000)
  - Liu (& Eichler) *et al*. Genome Res. 13 (2003)
    - 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)

Rates from growth model

- Average rates from model if *T = 4 By*

  $<r_S> \sim 0.25/site/By, \quad <r_D> \sim 0.50/Mb/My$

- About 7~8 time smaller than **recent** sequence divergence estimates
- 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* ~ 4 By

Bridging the two estimates

- Empirical rates $r_{S,D}$ for last ΔT~50 My are **terminal rates**
- Model rates $\langle r_{S,D} \rangle$ averaged over whole growth history, **hence**

$$\boldsymbol{\langle r_{S,D} \rangle < r_{S,D}}$$

- Given constant duplication rate r_D per length per unit time and constant average duplicated segment length λ, then genome grew exponentially. Fit to data gives

$$L(t) \sim 1\ (Mb)\ exp(t/0.5\ (By))$$

Remarks on $L(t) \sim 1\ (Mb)\ exp(t/0.5\ (By))$

- Our model can be reconciled with alignment based data on evolution
- HS genome 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 $t \sim$ *500My*, L ~ 1 Mb

Summary on genome data and growth model

- Genomes form a **universality class** defined by:
  - universal effective lengths
  - maximally self-similarity
  - Random correlation between words

- Genome-like sequence are generated by simple growth model characterized by:
  - Three universal parameters
  - maximally stochastic segmental duplication
  - Very early onset of duplication process

- For HS genome, model consistent with evolution rates extracted by sequence divergence methods

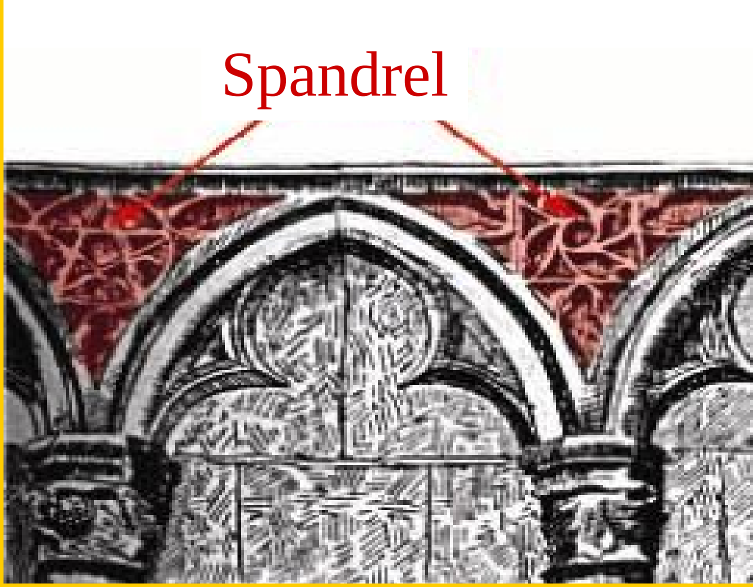
More phenomena explained by model

- Preponderance of **homologous genes** in all genomes
- Numerous **pseudogenes** in eukaryotic genomes (85% microbial genome is coded)
- Genome is full of **non-coding repeats**
- Large-scale genome "**rearrangments**"
- Rapid **rate of evolution** - random self-copying is an extremely efficient way for information accumulation; growth by random self-copying is likely the result of natural selection
- Many more ...

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)*



- Duplications to a genome are what the construction of arches, walls and ceilings are to a cathedral
- Codons are the spandrels and genes are décorations in the spandrels

Cross-disciplinary Similarities

Control theory	Economics	Biological cells	Genomes	Games
process variables	activities	phenotypic features	coding regions	board con-figurations
operating costs	activity costs	metabolic costs	information versus size	evaluation of board
objective function	profit	fitness	fitness	payoff
control policy	plan	reaction net	signal control net	strategy

Are genomes CAS's?

Characteristics of a *cas* (Holland)

A complex adaptive system, cas, is an evolving, perpetually novel set of interacting agents where

- **There is no universal competitor or global optimum.**
  There is no BEST genome (organism)

- **There is great diversity, as in a tropical forest, with many niches occupied by different kinds of agents.**
  There is great diversity in genomes

- **Innovation is a regular feature – equilibrium is rare and temporary**
  Genomes evolve continuously

- **Anticipations change the course of the system.**
  ??

- Genome is the system, genes and other codes are the agents
- Duplicated segments are the building blocks, site replacements (mutations) are the innovations
- Metaphor

Questions to be answered

- How to reconcile with or quantify effect of natural selection
- Can model be refined to differentiate coding and non-coding regions ?
- Can model be extended to describe the rise of genes and gene families, regulatory sequences, ...?
- Is model consistent with phylogeny?
- Can model say anything about the origin of life? (RNA world?)

‘Replicators’

- Early in the RNA world, RNA genes are thought to have been able to directly copy themselves (albeit imperfectly).

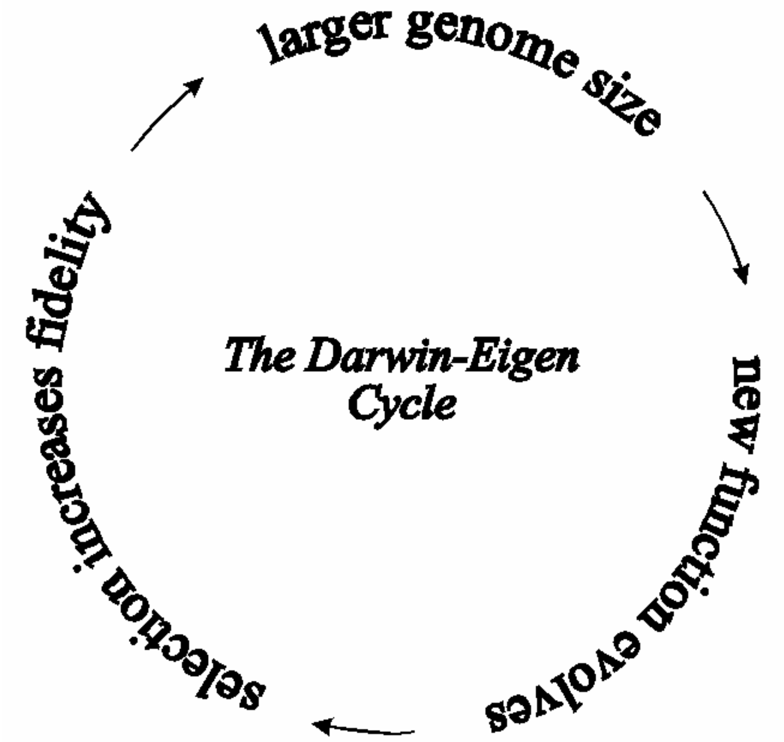
- Such RNAs are described as ‘replicators’.

The Eigen Limit: a paradox of prebiotic evolution

- *The amount of information that can be maintained in a genome is limited by the accuracy (fidelity) of replication.*

A. Poole, The RNA world & LUCA

Avoiding Catch-22



- Imagine the best replicator is shorter than max. length dictated by error threshold
- A longer mutant with higher copying fidelity can emerge, which allows new max. length
- Longer sequences are sequentially possible, and stepwise increase in replication fidelity occurs.

Scheuring (2000) Selection 1:135
Poole et al. (1999) BioEssays 21:880-889

More fundamental problems

- Cause for variation in base composition – why is base composition different from organism to organism but (almost uniform in a genome?
- How to reconcile universal growth model with apparent genome specific substitution rate?
- HS genome is still growing (our luck) but microbial genomes must gained its current size 2-3 billion years ago. How do microbial genomes maintain “universal” stats property under effect of constant mutation but with no growth?
- Can our model be COMPLETELY WRONG?

Some recent papers on segmental duplication and evolution

- Science, Vol 297, Issue 5583, 1003-1007 , 9 August 2002
  Recent Segmental Duplications in the Human Genome
  Jeffrey A. Bailey, et al. (Eichler group)
  http://www.sciencemag.org/cgi/content/full/297/5583/1003
- Genome Res. 2003 March 1; 13(3): 358-368.
  Analysis of Primate Genomic Variation Reveals a Repeat-Driven Expansion of the Human Genome. Ge Liu, et al (Eichler group)
  http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=430288
- Science, Vol 297, Issue 5583, 945-947 , 9 August 2002
  Gene Duplication and Evolution. Michael Lynch
  http://www.sciencemag.org/cgi/content/full/297/5583/945
- Science, Vol 290, Issue 5494, 1151-1155 , 10 November 2000
  The Evolutionary Fate and Consequences of Duplicate Genes
  Michael Lynch and John S. Conery
  http://www.sciencemag.org/cgi/content/full/290/5494/1151

‘Accessible’ reading on LUCA and the RNA world (Anthony Poole)

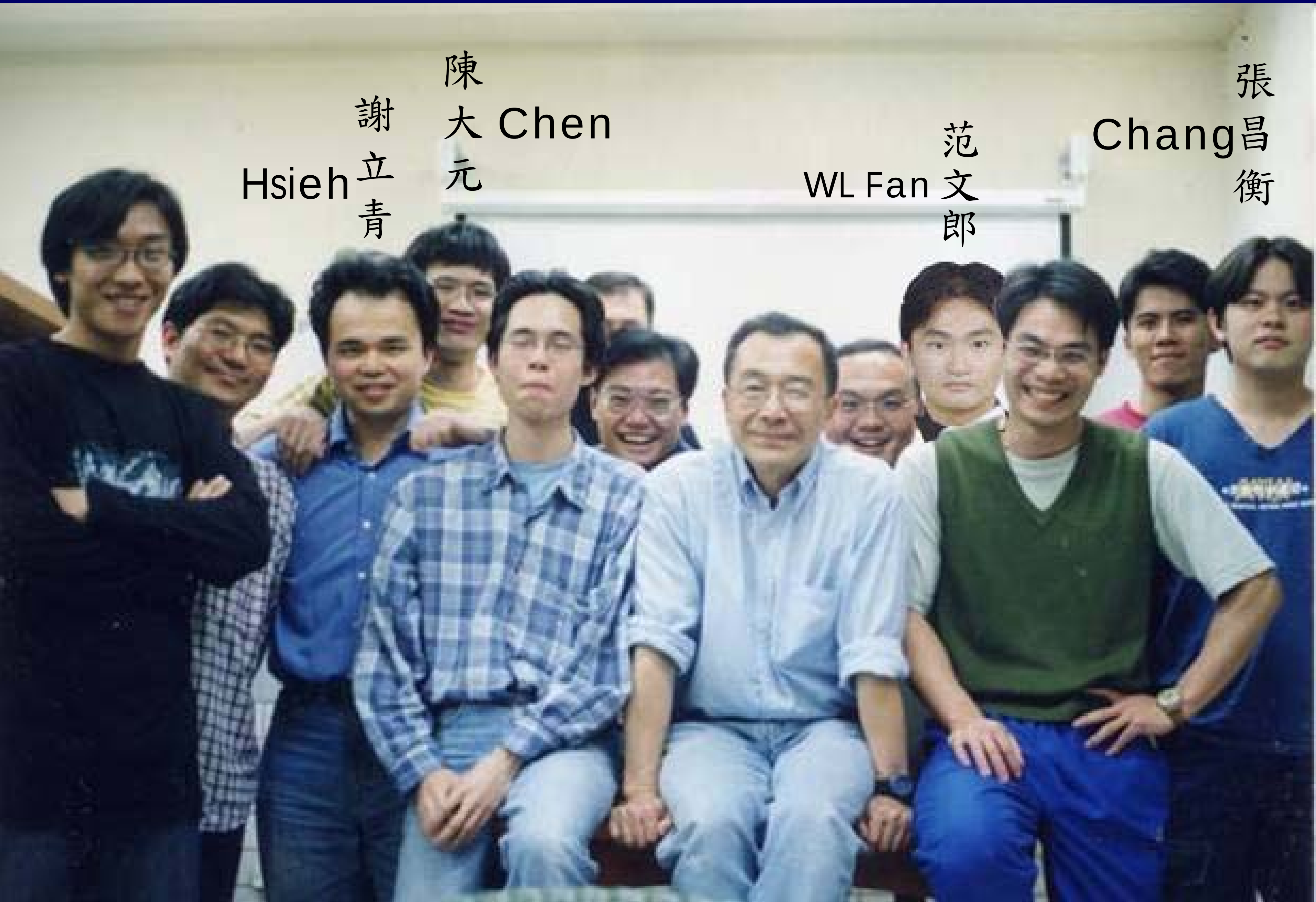
- http://www.actionbioscience.org/newfrontiers/

- Ridley (2000) The search for LUCA.
  Natural History November pp. 82-85.

- Morton (1999) Making life simple.
  New Scientist 16th January pp. 34-37.

- Pennisi (1998) Direct descendents from an RNA world.
  Science 1st May, 282:673.

Computation Biology Laboratory (2003)



Our papers are found at Google: HC Lee

Thank you!