The Blind Self-Copier –

Entropy, universality, inflation & spontaneous symmetry breaking in the growth of genomes

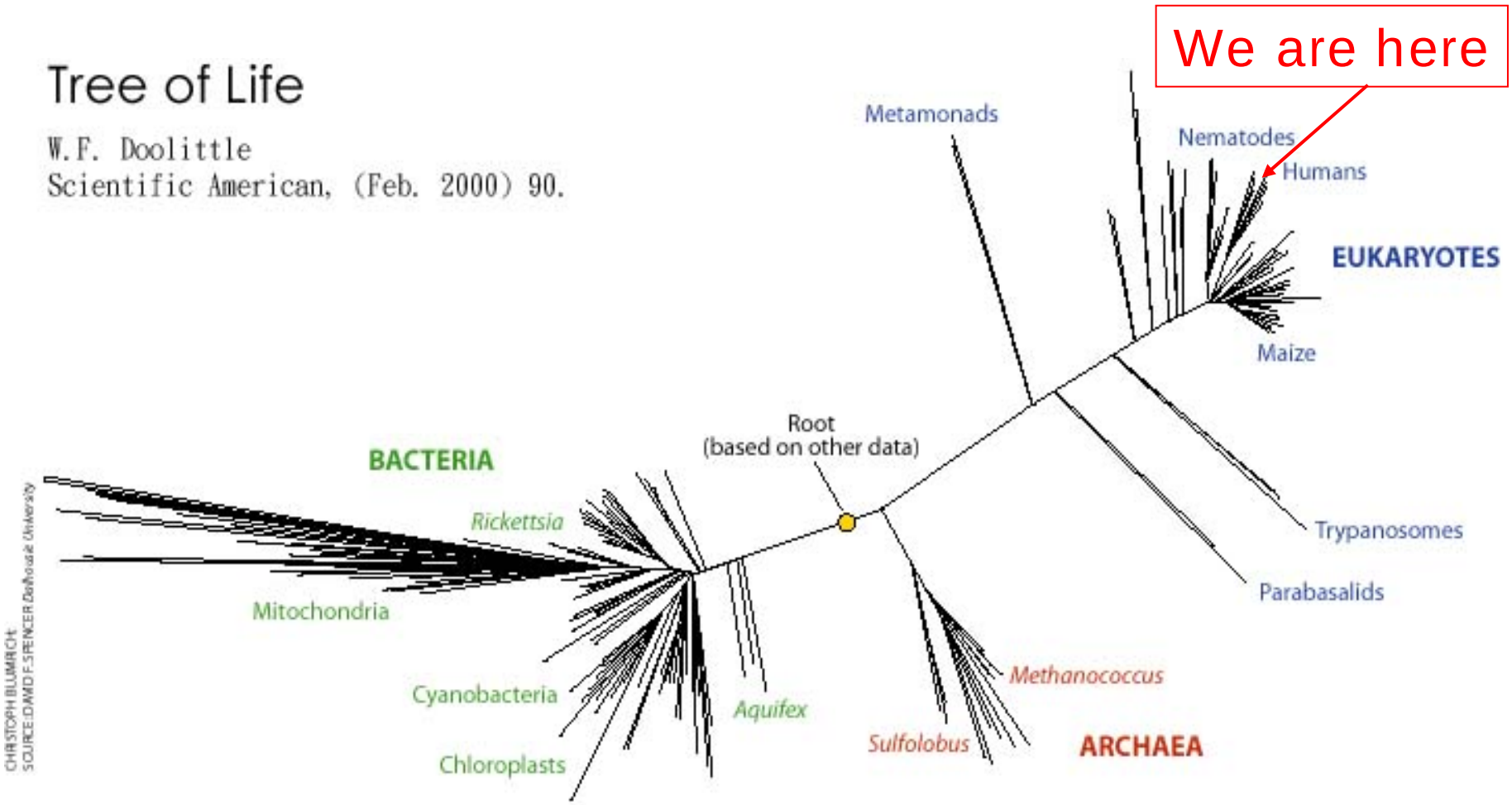
19th Spring School on Particles & Fields
Chung Yuan Christian University
Chungli, Taiwan
2006 April 2

HC Lee
Computational Biology Lab
Dept. Physics, Inst. Biophysics & Inst. Systems Biology
National Central University

Some concepts to be discussed

- Examining the growth and evolution of genomes, we consider
  - Randomness and order
  - Second law of thermodynamics
  - Distribution and its variance
  - Entropy and (Shannon) information
  - Diversity and universality
  - Complexity and self-similarity
  - “Inflation”
  - Spontaneous symmetry breaking
  - Neutral evolution and natural selection
- and arrive at a hypothesis and model for genome growth

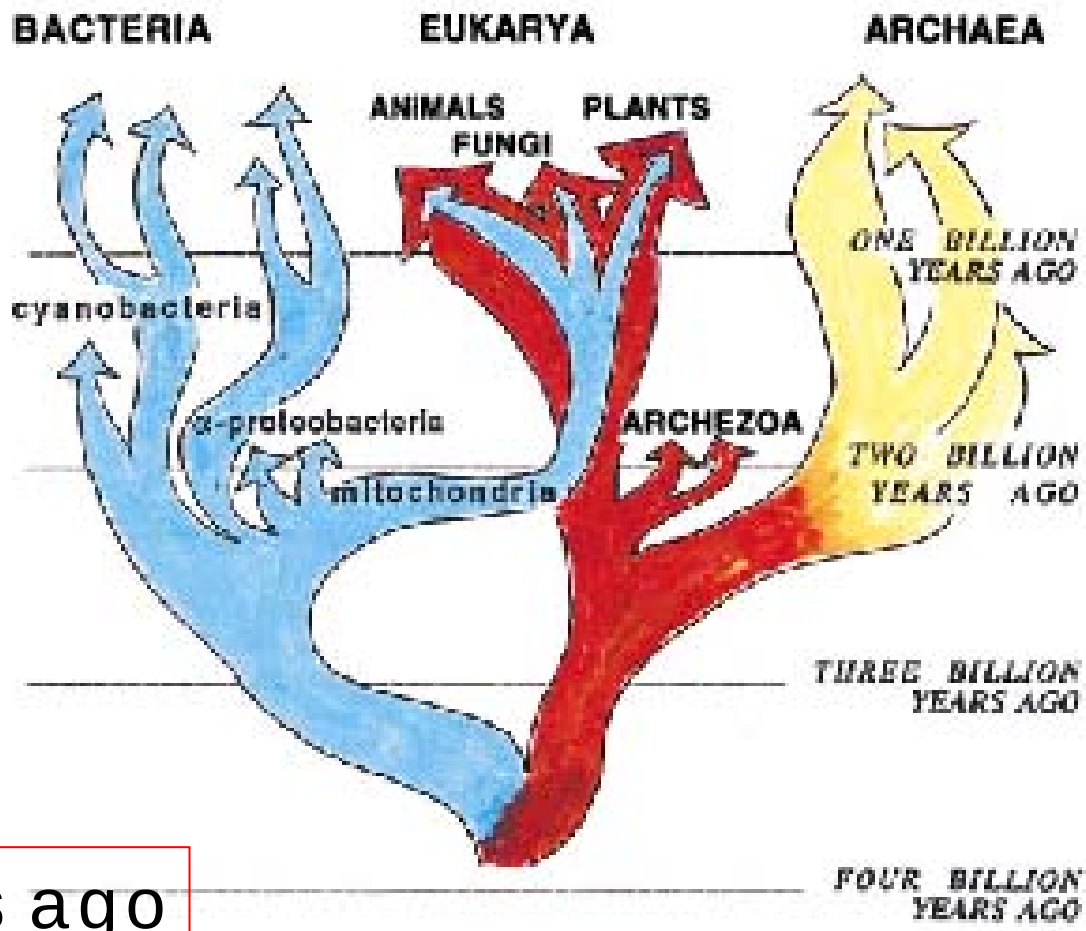
Life is highly diverse and complex



And it took a long time to get here

now

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



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 chromo-some of Homo sapien

(from the 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 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
```

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

Growth of genome – universality & inflation

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
  - $L \sim \tau/(CV)^2$

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

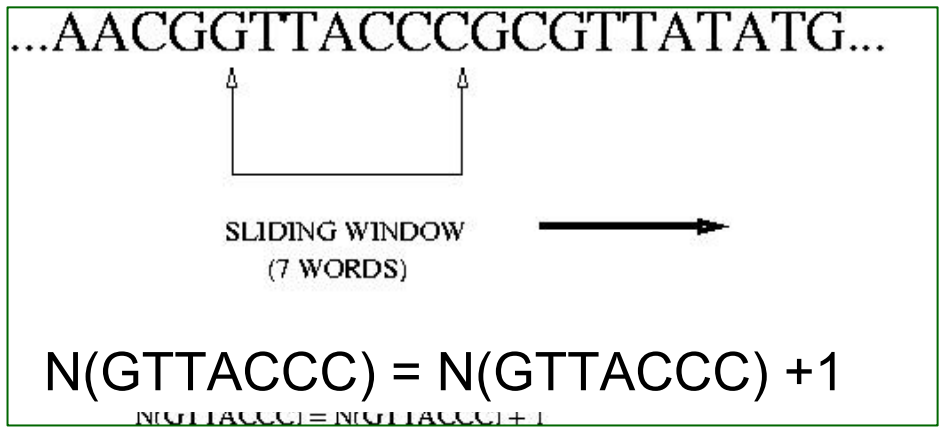
$$R = L^{-1} \Sigma_i f_i \ln(f_i/f_{ave}) = (CV)^2/2 + \ldots$$

- Shannon information and coefficient of variation are equivalent measures

(Note: technical detail re biased base composition important)

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



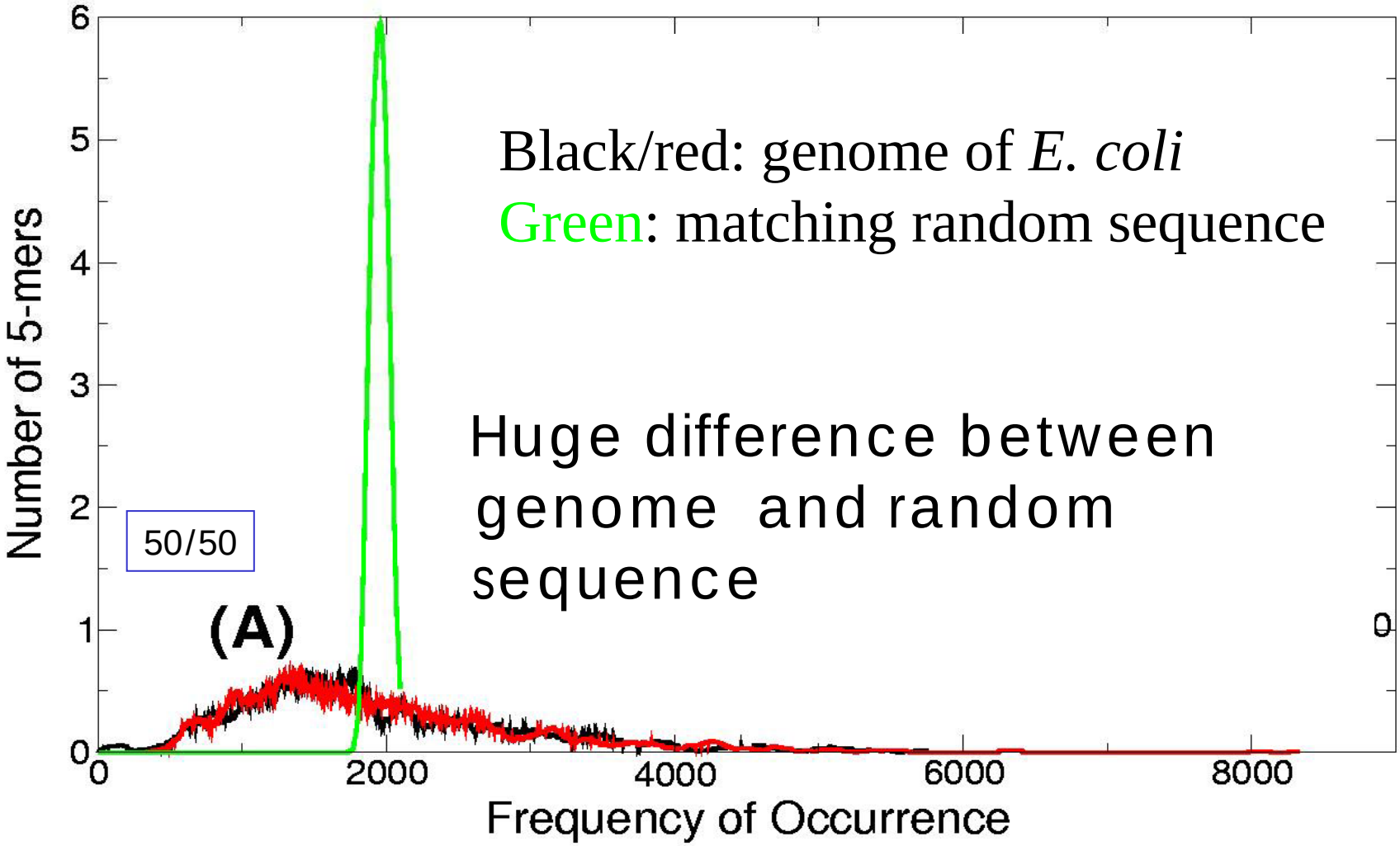
For genomes: frequency distribution of occurrence of words

- Events: word occurrence;
- Type of events τ: types of words = 4^k;
- Distribution: distribution of frequency of occurrence
- Given τ and CV, define effective length

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

  - The randomness (or CV) of the measured sequence is the same as that of a random sequence of length L_{eff}

A test case: genome of *E. coli* *4.6 Mbp, 25% each of A,C,G,T*

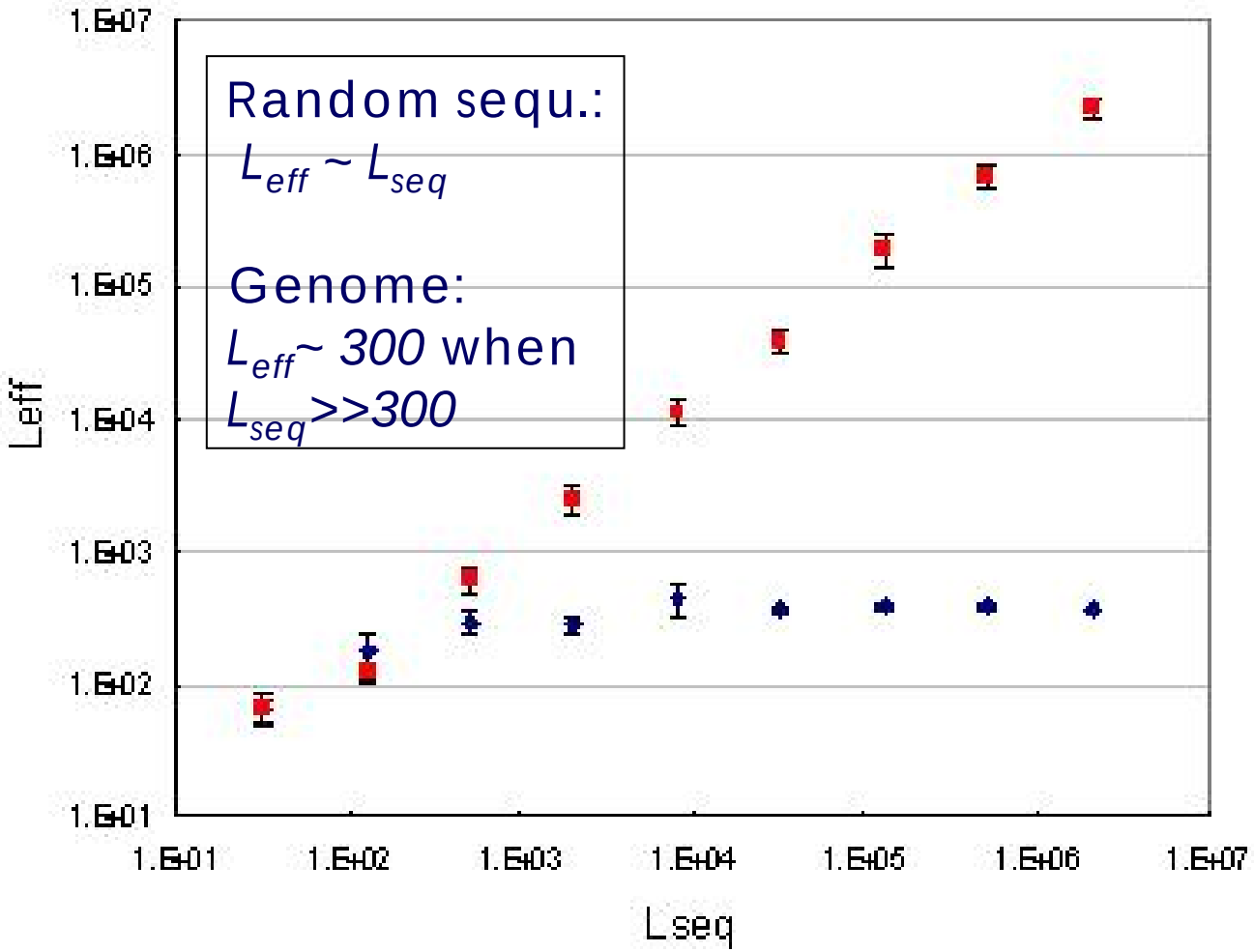


L_{eff} does not grow with sequence length, but saturates

- random sequ.
- *E. coli*

Each datum: Average L_{eff} of 30 randomly selected segments of length L_{seq}.

E. coli, 4.6 Mbp, 50% A+T, 50% C+G

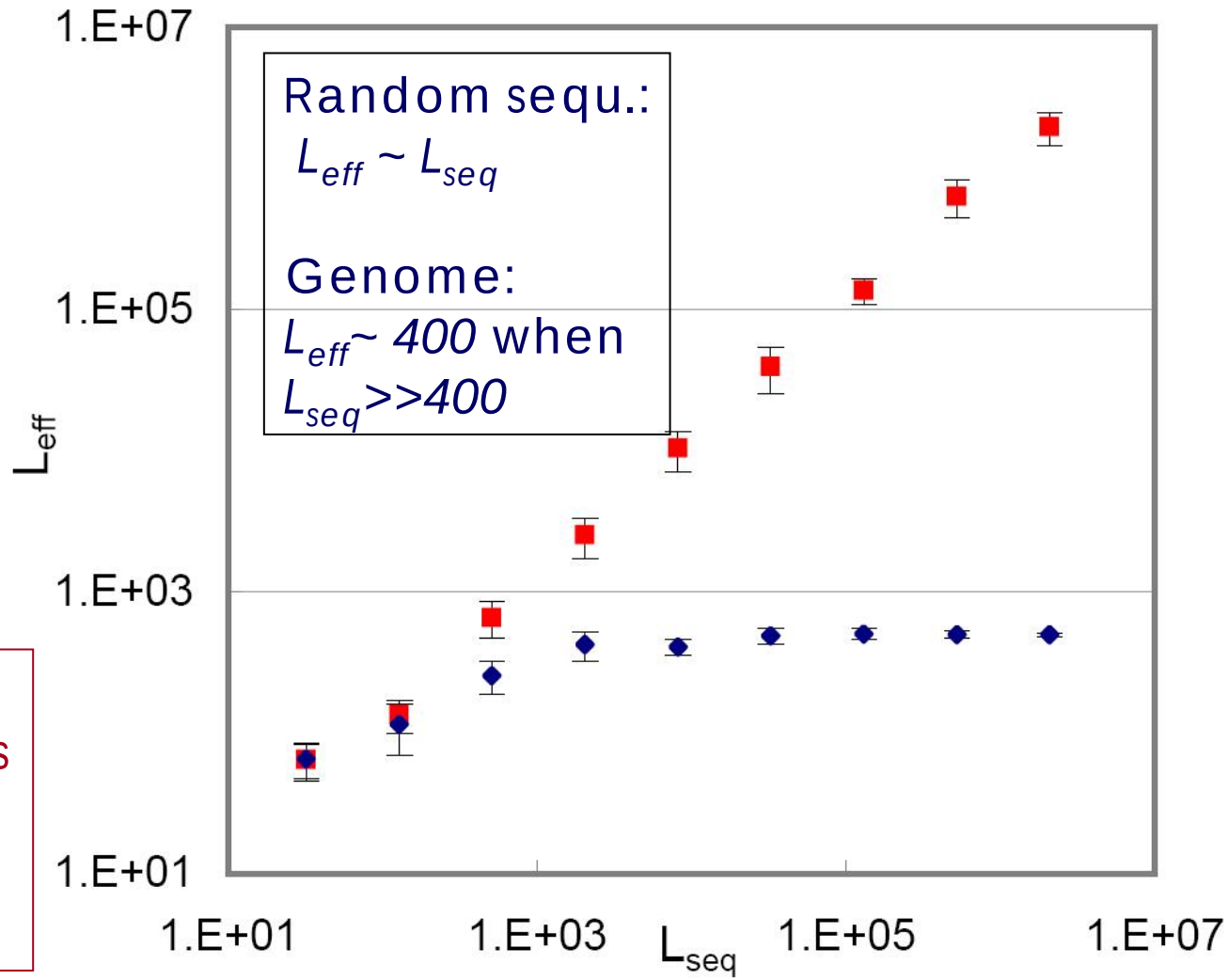


L_{eff} does not grow with sequence length (cont'd)

A. thaliana Chr. I, 30 Mbp, 64% A+T, 36% C+G

- random sequ.
- *A. thaliana*

Each datum: Average L_{eff} of 30 randomly selected segments of length L_{seq}.



For 2-letter words genomic sequences has the randomness of 300~400 base random sequence

Two big surprises from complete genomes

- 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

Small L_{eff} implies large Shannon info R

Genome: *Pyrobaculum aerophilum*, 2.22 Mbp, p=0.5

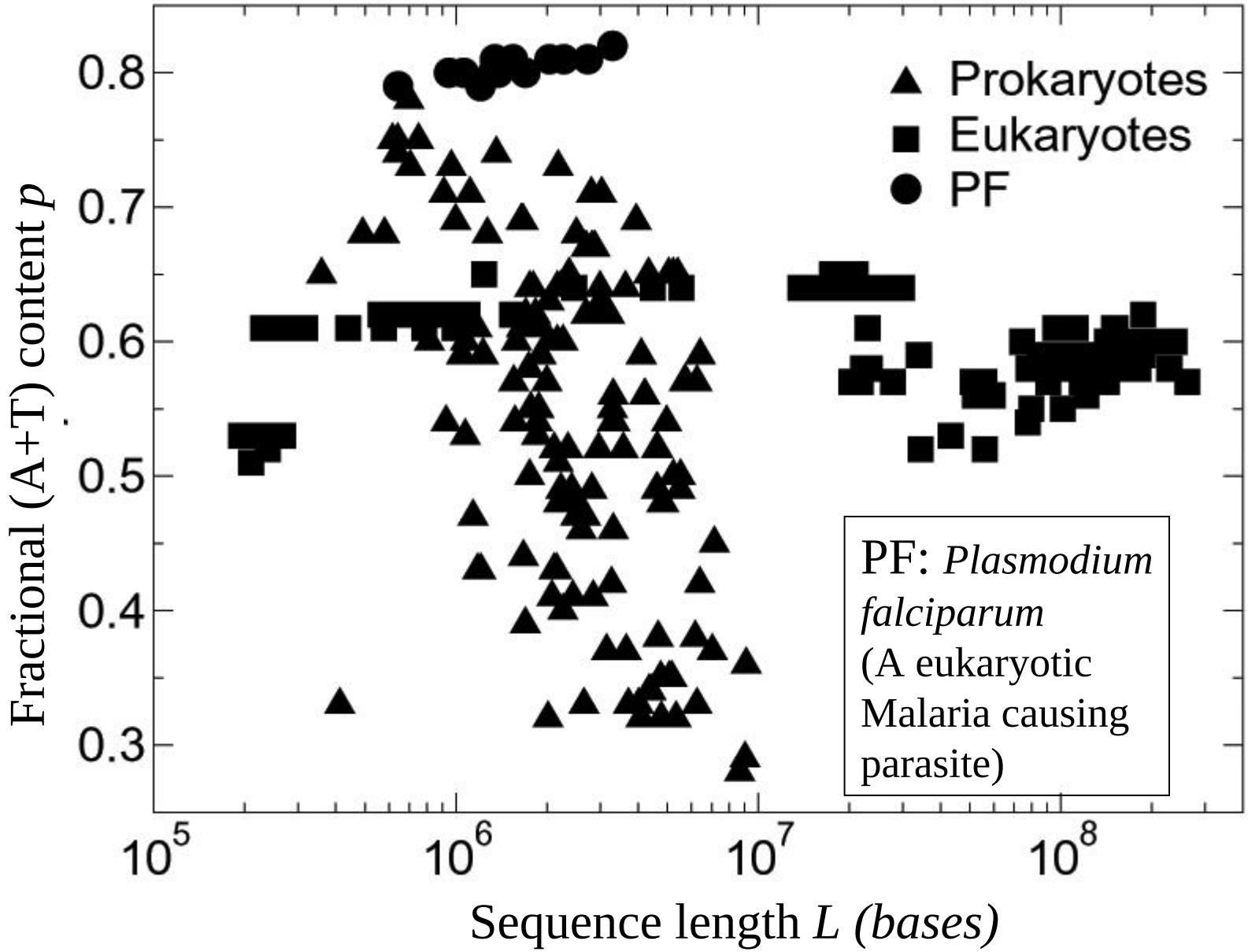
	Random sequence			Genome sequence		
k	H	R	R_{ex}	H	R	R_{gen}/R_{ran}
2	3.9999	5.90 E-6	5.77 E-6	3.973	2.66 E-2	**4500**
3	5.9999	3.72 E-5	3.46 E-5	5.933	6.65 E-2	**1922**
4	7.9999	1.72 E-4	1.62 E-4	7.881	1.18 E-1	**728**
5	9.9993	7.26 E-4	7.53 E-4	9.821	1.79 E-1	**246**
6	11.999	2.94 E-3	2.90 E-3	11.75	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	19.048	5.92 E-1	5.24 E-1	18.59	5.92 E-1	1.4

Note: R is a useful quantity; H is not

Now look at all complete genomes in *GenBank*

- ALL (278) complete sequences from **GenBank** at the time of download (Nov. 2004)
- Include 165 complete prokaryotic (原核) genomes and 113 complete eukaryotic (真核) complete chromosomes
- For each sequence compute $L_{eff}(k)$ for k= 2 to 10 (or to k_{max} such that sequence length > 4^{kmax})
  - Each k has about 278 pieces of data
  - All told about 2000 pieces of data

Complete Genomes are diverse



Compute L/L_{eff}
Large ratio implies less randomness, more order, more "information"

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

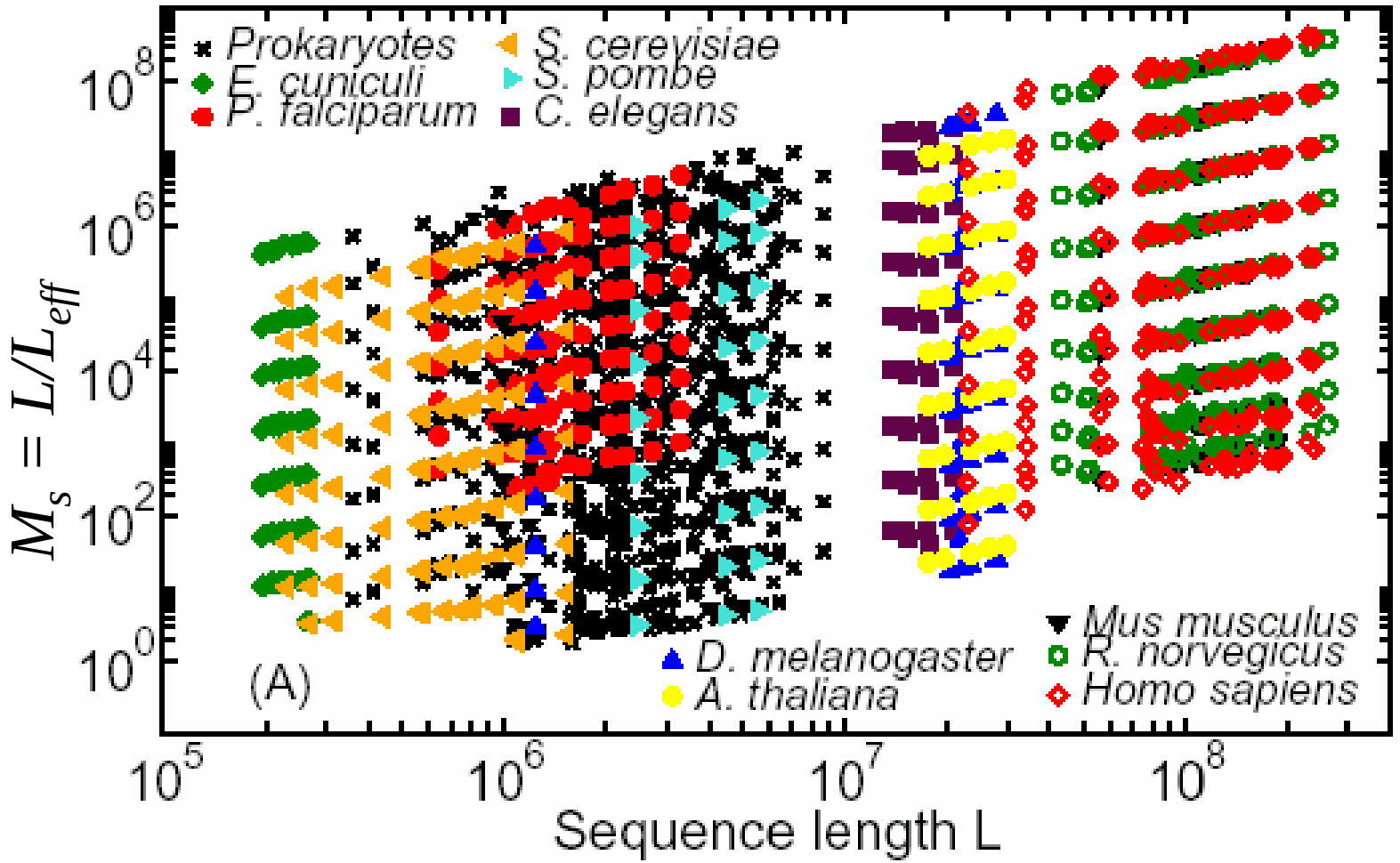
$$(CV_{genome})^2 = \tau/L_{eff}$$

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

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

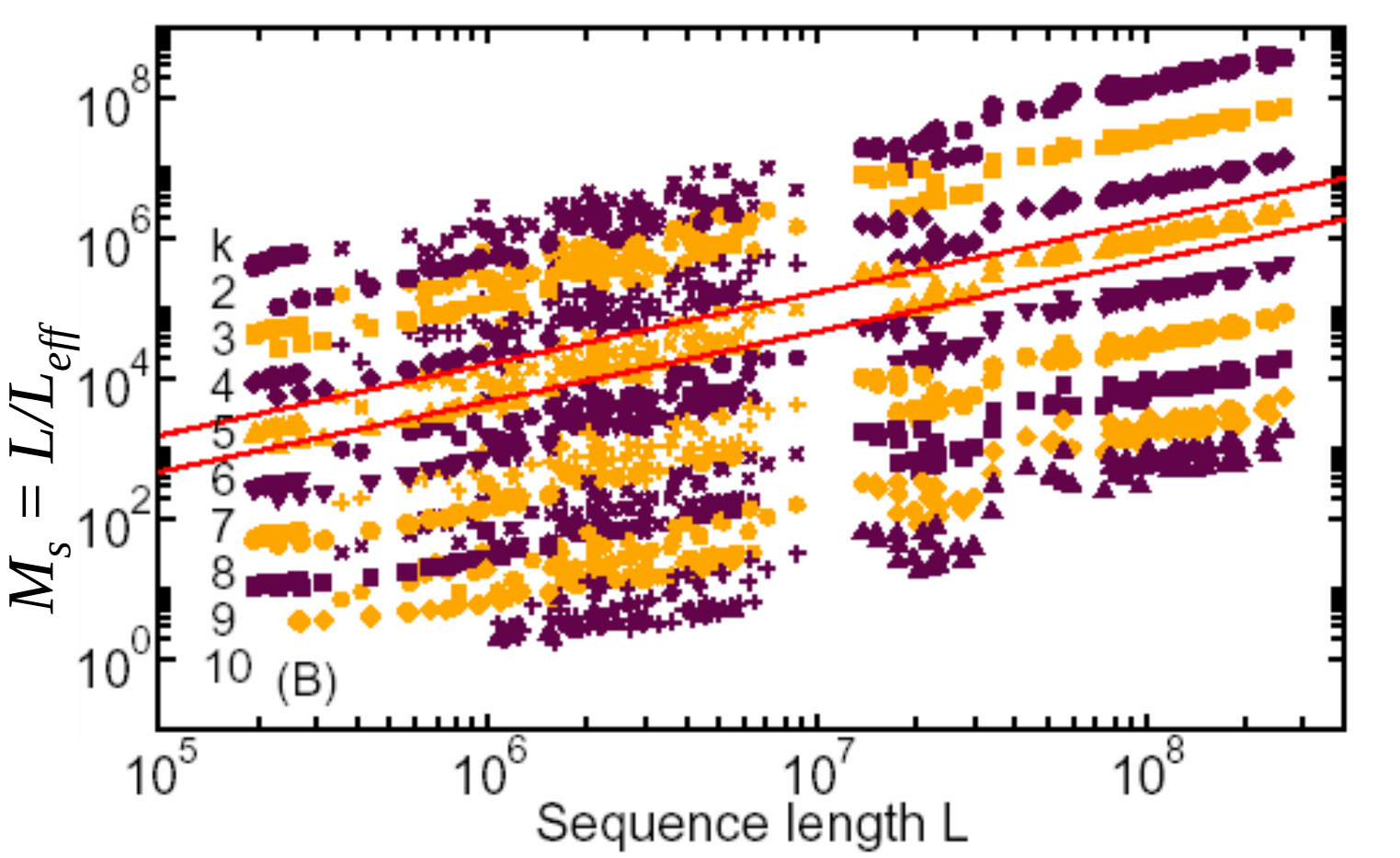
Note: technical complications when p not equal to 0.5

Results: color coded by organisms



Each point from one *k*-spectrum of one sequence; ~2000 data points. Black crosses are microbials. Data shifted by factor 2^{10-k}.

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



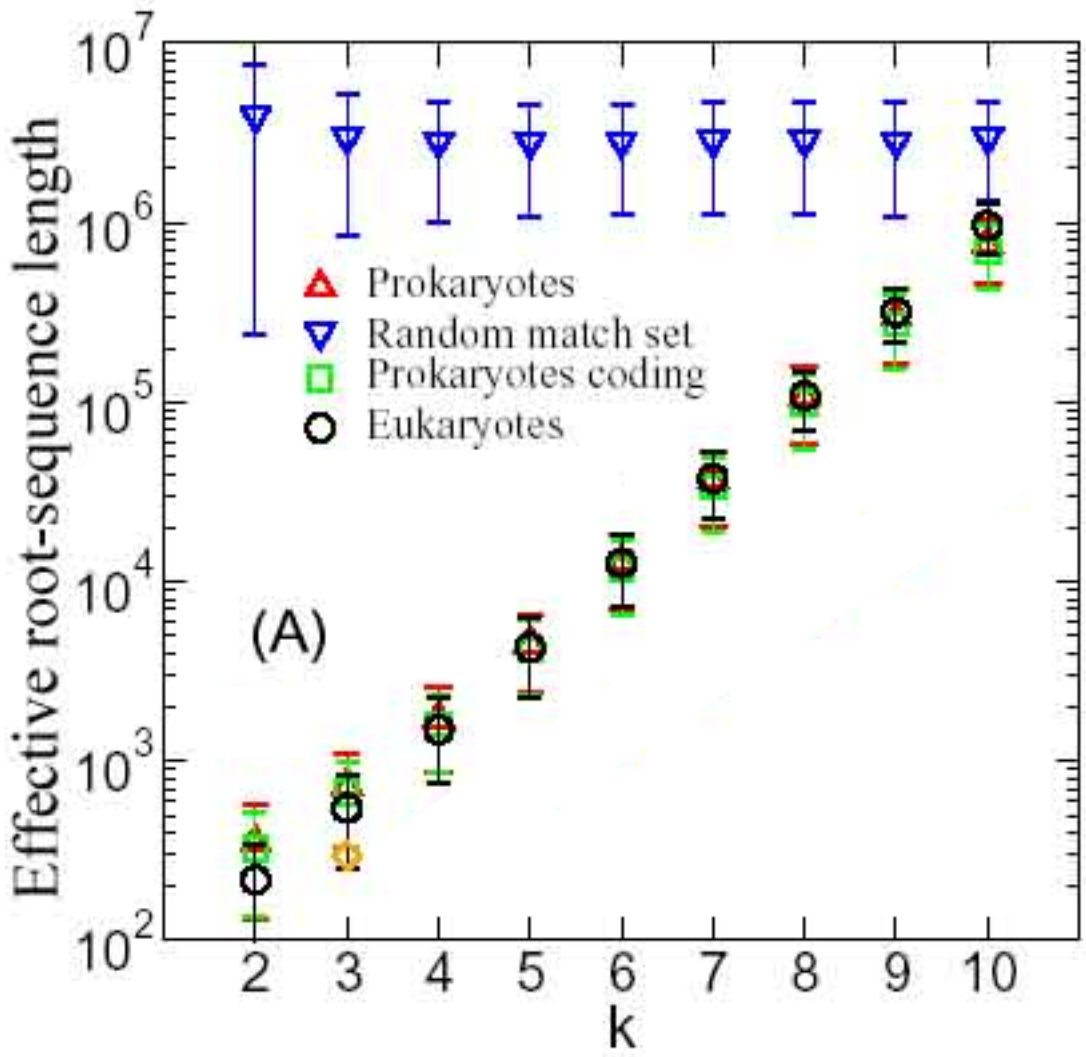
Data from 14 *Plasmodium* chromosomes excluded; ~2000 data points. For each *k*, 268 data points form a narrow $M_\sigma \sim L$ "*k* band".

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 given be two parameters.
  a = 0.398+-0.038
  B = 1.61+- 0.11
- Defines a **universal class**
- Mild exception: Plasmodium



Black: genome data; green: artificial

Summary of genome data

- Universality class – for fixed word length k, L_{eff} is (approximately) the same for all genomes and independent of genome length
  - 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?

Duplication reduces randomness and L_{eff}, and increases order

- 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

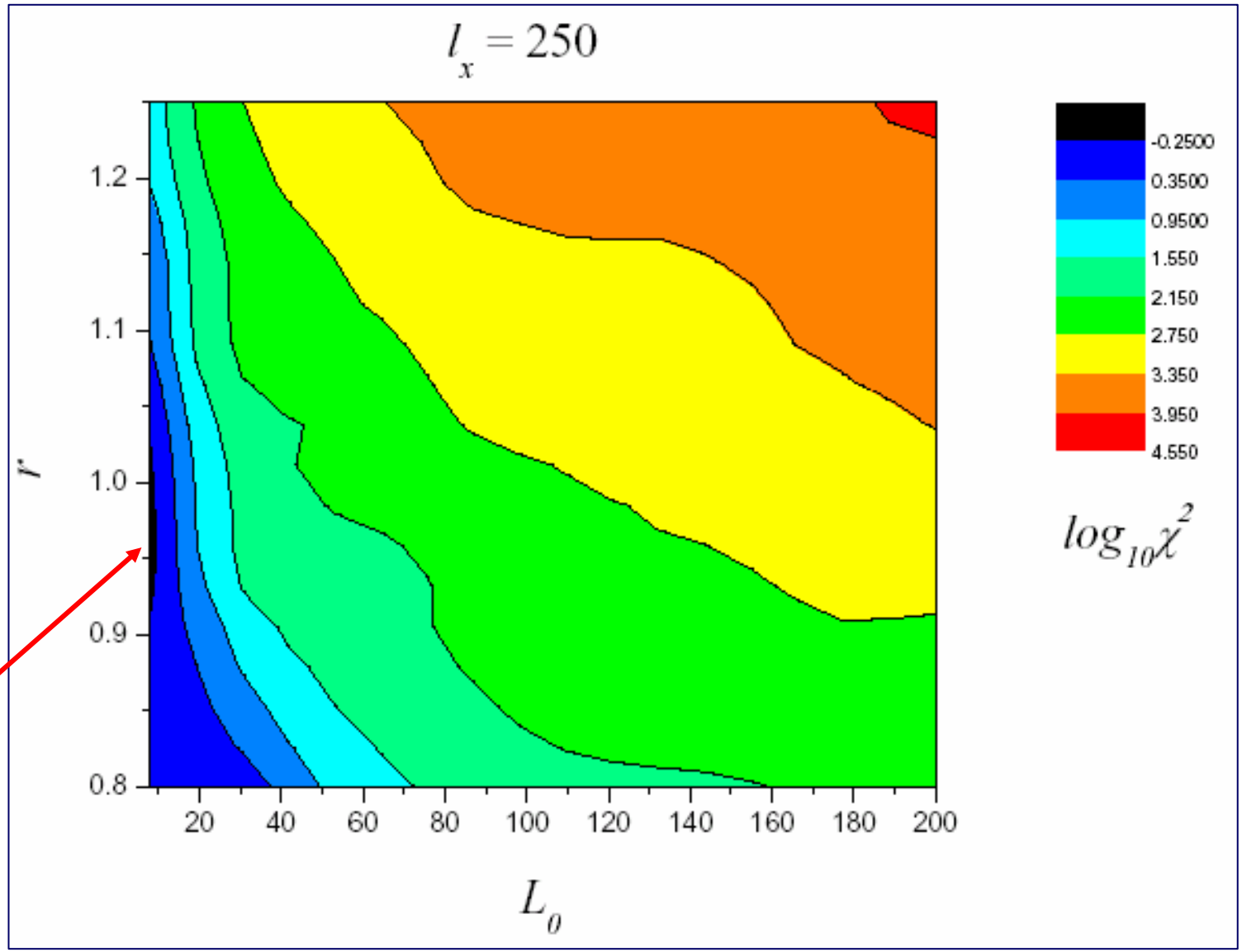
A model: at a universal initial length, genomes grew (and diverged) by maximally stochastic segmental duplication

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

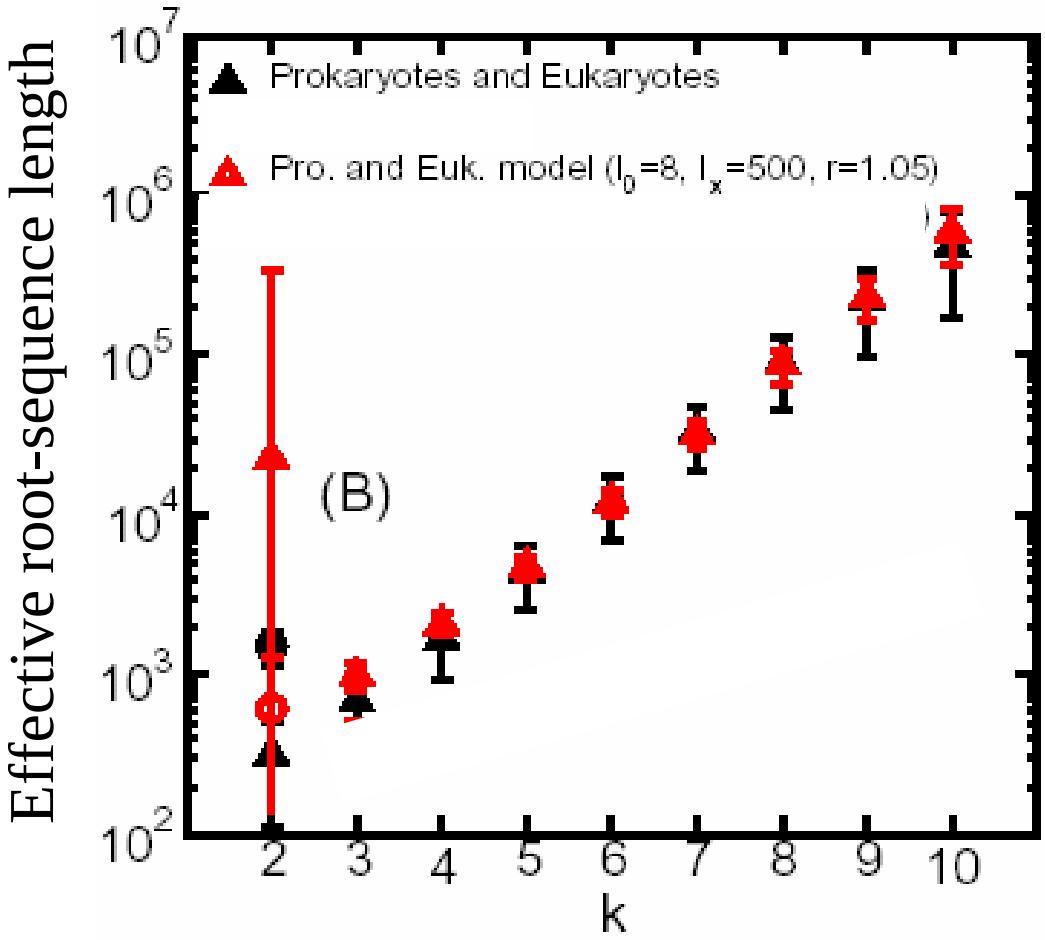
Self copying – strategy for retaining and multiple usage of hard-to-come-by coded sequences (i.e. genes)

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

Model parameter search: favors very small L_0



Model with three universal parameters – successfully generates universal L_{eff}



Red symbols are from 278 genome matching model sequences

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:
  - Genomes are Blind Self-Copiers – growth by **maximally stochastic segmental duplication**
  - Very early onset of duplication process
  - Model w/ three universal parameters successful

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

(Self)-copying: paradigm for rapid information acquisition, and for winning the "survival-of-the-fittest" war

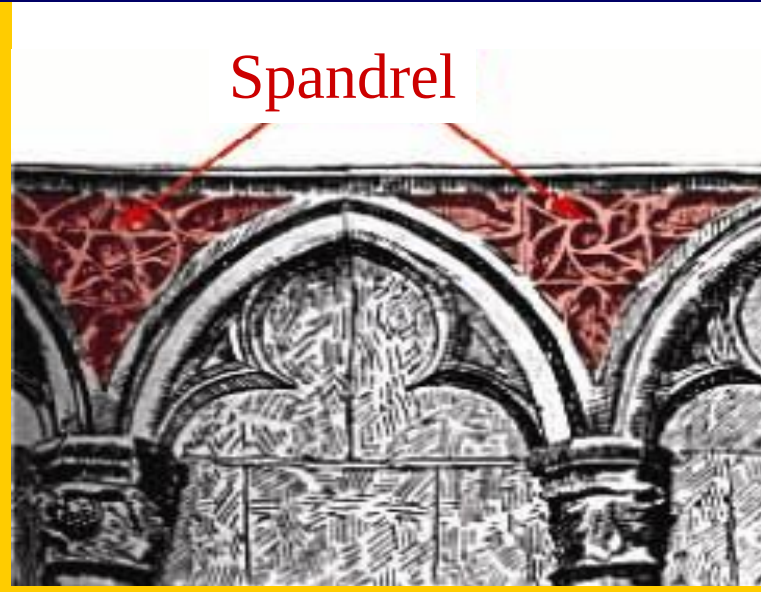
- Invention is hard, copying is easy
- Large system requires protocol (standard codes), which MUST be copied
  - Everyone's first C-program code is probably a copy
  - Subsequent codes made by copy + alteration
  - Same goes with TeX/LaTeX codes, and others
  - If not self-copy, then plagiarize
- Life forms (not our ancestor) that did not invent copying machinery very early on could not have survived

Many biological 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
- The **RNA world** & 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

Big mysteries remain

- Why is there (almost) no difference in the L_{eff} of coding and non-coding regions?
  - Coding regions are supposed to be protected by *natural selection* and non-coding regions are not

- Bacterial genomes stopped growing ~2 billion years ago, why have the traces of stochastic duplication not been eroded by random mutation?

Was there inflation in early genome growth?

- Our model works only if most mutations take place *after* growth is completed
  - If duplication and random mutation happened concurrently, L_{eff} would still grow with total genome length and not stay constant (as observed)
- That is, our model describes a scenario of early inflationary growth without mutation. Mutations came late in the life of genomes.
  - Explains the intra-uniformity of genomes
  - Idea biologically exotic, needs to be further tested
  - The macronucleus of *Tetrehymena* (纖毛蟲) contains ~50 copies of each gene, except for rRNA genes, which exist in ~10,000 copies.

Emergence & spontaneous symmetry breaking in evolution of genome

Emergence in evolution

- Major steps in evolution are phenomena of emergence: *from nothing to something*
  - Biomolecules
  - Membranes
  - Genome
  - Primitive copying machinery
  - RNA genes
  - DNA genes
  - Codons and machinery for the Central dogma
  - Reproduction schemes
  - Many more ...

Spontaneous Symmetry Breaking

- Emergence is SBB: *from nothing to something*

Symmetry preserved	Symmetry broken
Uniformity	Non-uniformity
Everything is the same	Something is different
(Nothing)	(Something)

- How does emergence happen?
  - Here, simulate it with *Cellular Automata*

DNA uptake signal sequence

- The DNA of some naturally "competent" species of bacteria contains a large number of evenly distributed copies of a perfectly conserved short sequence.

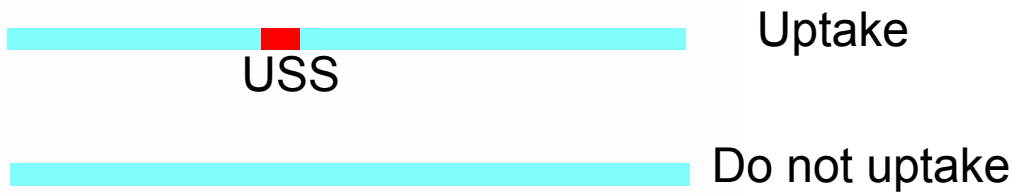
- This highly overrepresented sequence is believed to be an uptake signal sequence (USS) that helps bacteria to take up DNA selectively from (dead) members of their own species.

Competent bacteria have highly over-represented USSs

- *Haemophilus influenzae* has 1747(USS)/1.83 Mb, or ~1/kb, expected frequency is ~$5x10^{-3}$/kb
- *Neisseria gonorrhoeae* and *N. meningitidis*: 1891/2.18 Mb ~ 0.9/kb
- *Pasteurella multocida*: 927/2.26 Mb~0.4/kb
- *Actinobacillus actinomycetemcomitans*: 1760/4.50 Mb ~ 0.4/kb

Some USS issues

- USSs are evenly distributed over host genome
- Host organism preferentially uptakes DNA with USS



  - That is, they "eat" the DNA fragments of their dead relatives
  - Uptaken DNA digested as food or used for replacement of host genome
- Also known: USS bearing DNA uptaken by unrelated species

USSs are embedded in genome in a way that minimizes cost

- More USS per base in non-coding regions than in coding regions
- When embedded in a gene, USS preferably resides in less conserved areas
- Embedment of USS is gene slightly reduces conservation of embedding site

Cost & benefit issues

- Benefit
  - DNA has nutritional value
  - Homologous DNA (those w/ USS) for repair and/or recombination
- Cost
  - Takes up significant (~3% in *H. influenzae*) part of genome
  - Interferes with coding (hence most USS in non-coding or, if in coding region, then preferably in segments of relatively low conservation)

USS favors non-coding regions

Sequence class[1]	Length(bps)		Number of USS		Number of shared-USS
(1) non-coding	191,088	10.44%	496	33.72%	24*
(2) rRNAs, tRNAs and Structural RNAs	31,868	1.74%	0	0%	---
(3) ribosomal proteins	23,910	1.31%	3	0.2%	---
(4) proteins of assigned function	1,084,671	59.27%	633	43.03%	17*
(5) putative and hypothetical proteins	501,810	27.42%	363	24.68%	7*
(6) overlap genes[2]	3,207	---	0	---	---
Total	1,830,140	100.18%#	1471	101.63%$	---

1 Only the 100% conserved 9-mer part of the USS is considered. Both USS+ and USS- in the coding regions are counted.

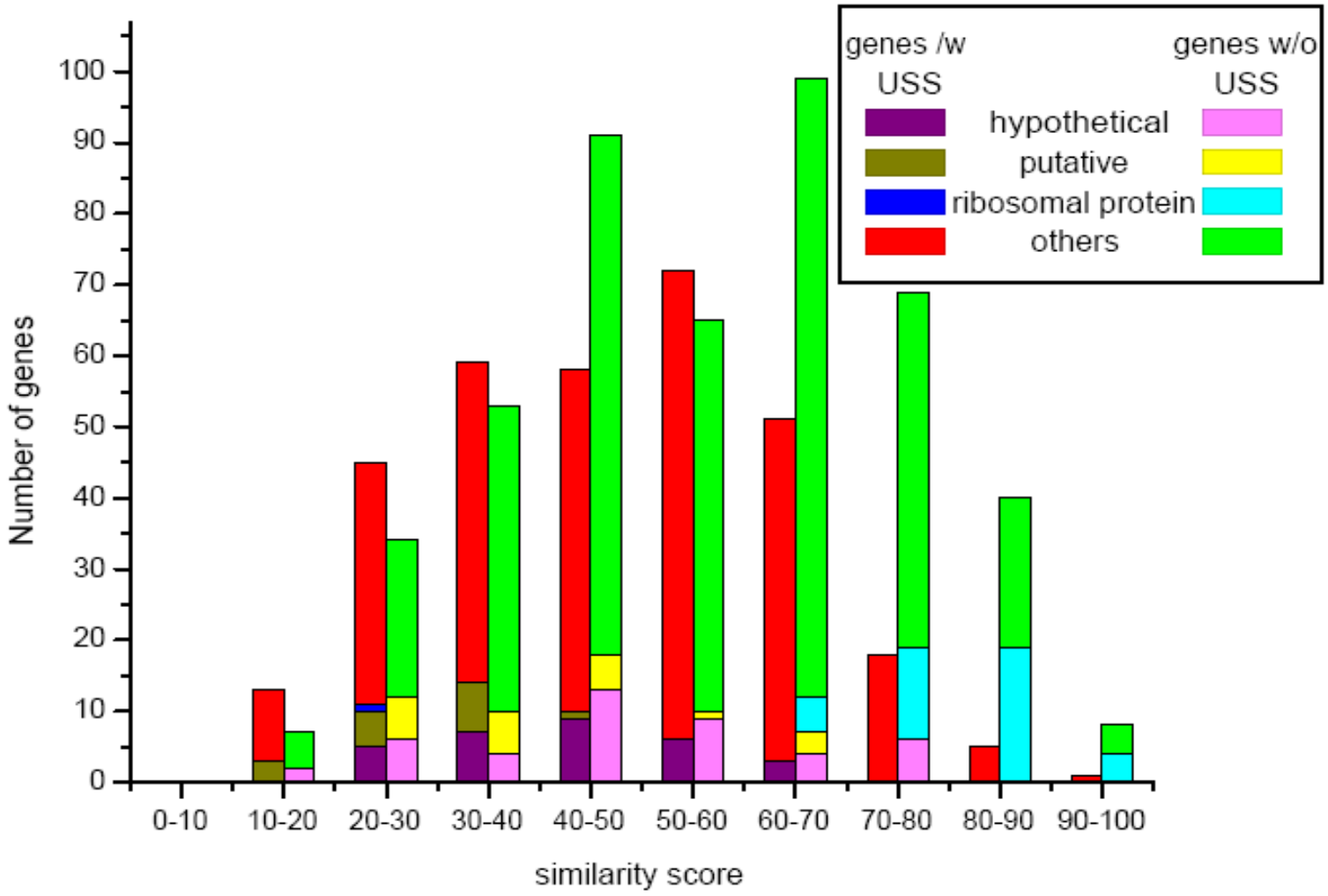
2 Including cases where both genes are in the same strand and cases where one gene in plus strand and the other in minus strand.

* 17 USSs straddle on both a class (1) and a class (4) region; 7 USSs straddle on both a class (1) and a class (5) region.

Exceeds 100% owing to overlaps.

$ Exceeds 100% owing to shared-USSs.

USS favors non-coding regions



Segmental scores in genes

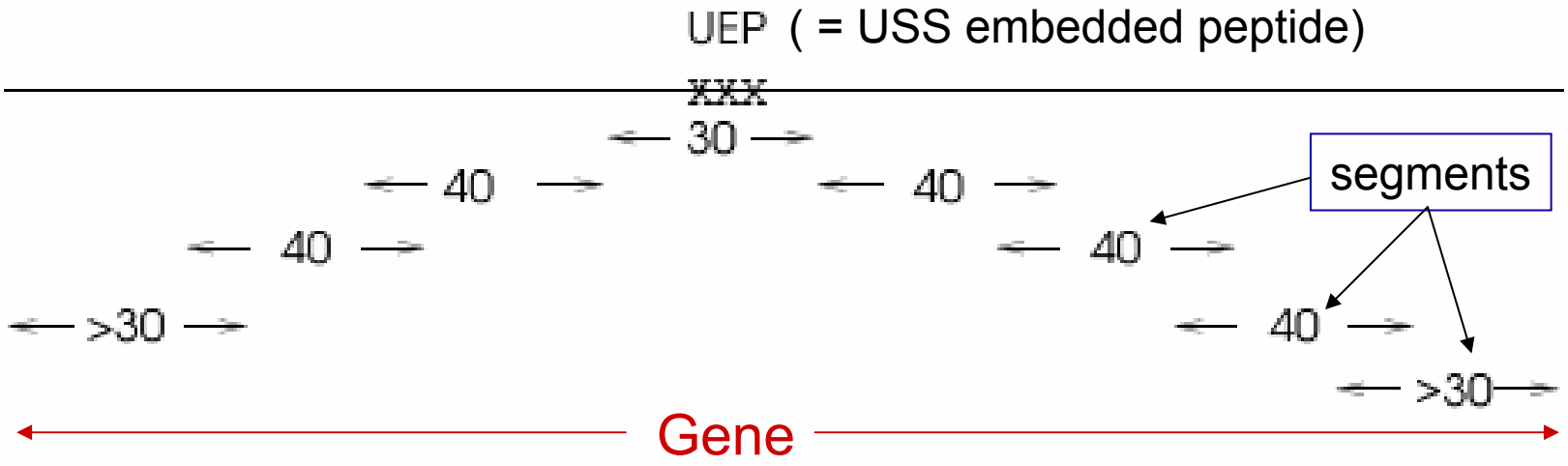
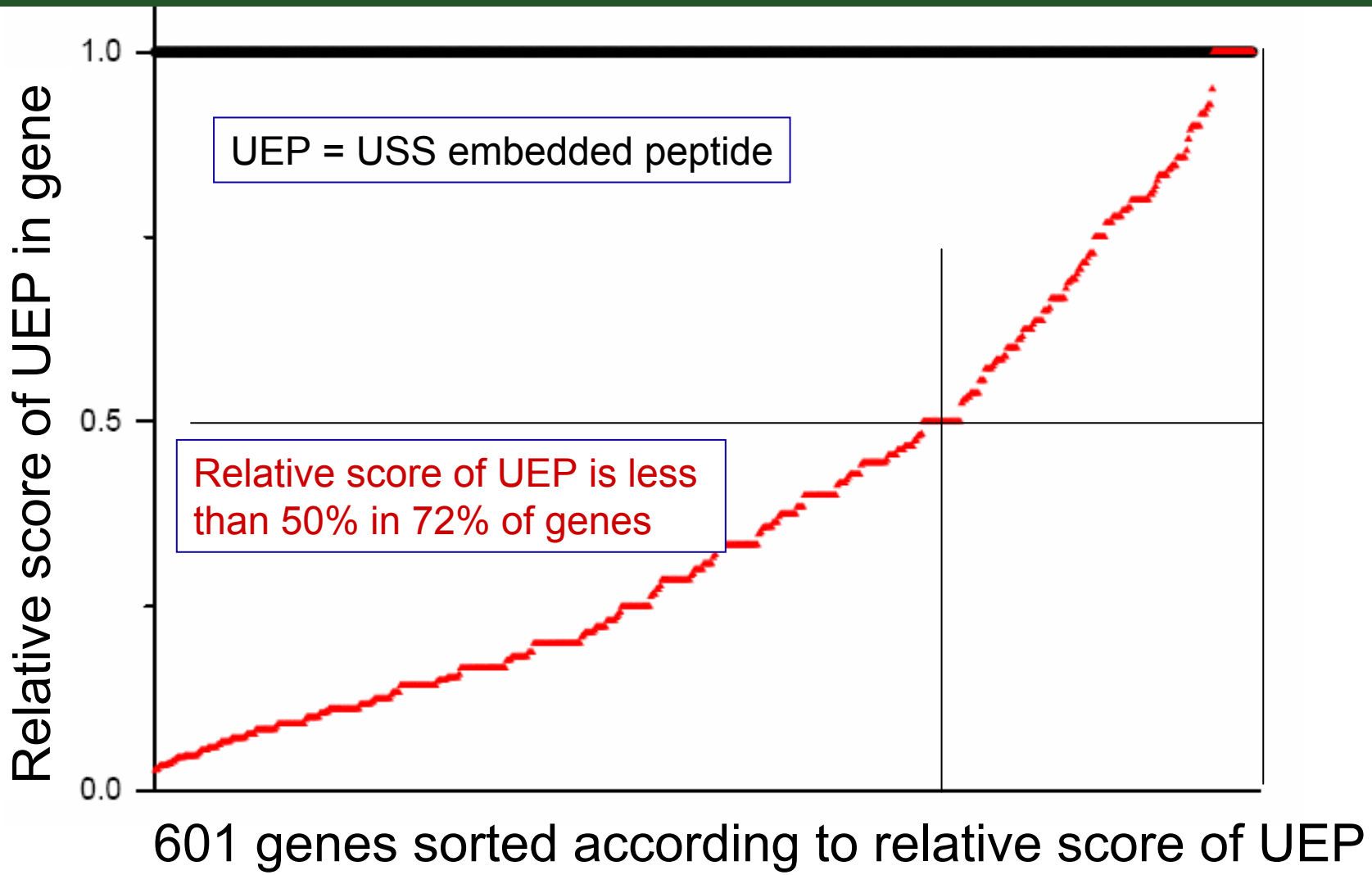


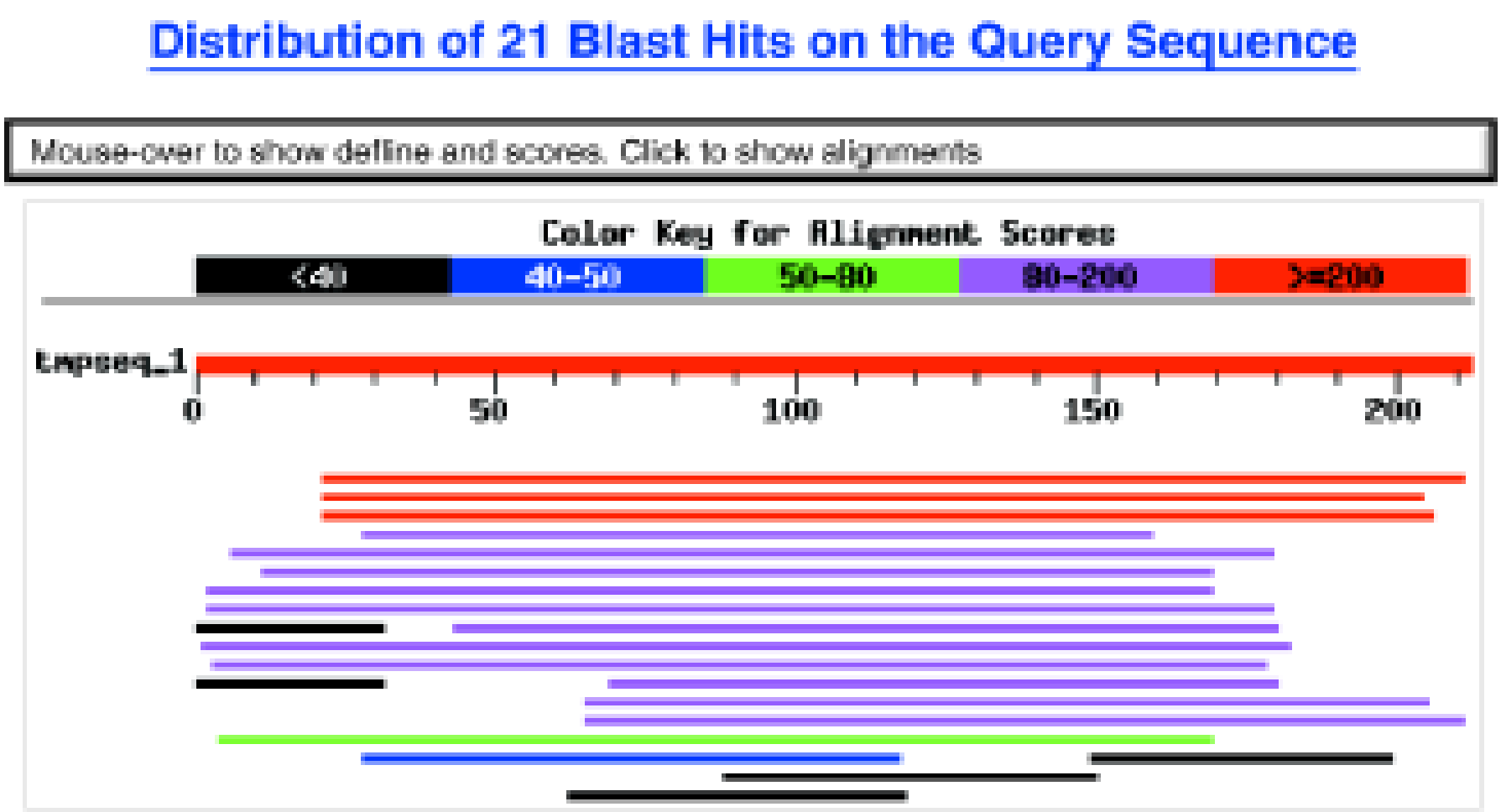
Figure 1: Segmentation of a protein sequence. XXX is the position of the UEP in the query, or of the corresponding peptides in a match.

Percentile score	Mean Score of all Segments	Number of all Segments	Mean Score of UEP's Segments	Number of UEP's Segments
0-1	44.83 +- 20.97	8,321	32.02 +- 19.94	601
0-0.25	44.47 +- 20.79	4012	17.39 +- 12.51	271
0.25-0.5	45.26 +- 20.54	2269	35.72 +- 11.82	170
0.5-0.75	44.62 +- 22.89	1086	45.54 +- 15.84	79
0.75-1	45.55 +- 20.45	954	60.02 +- 13.50	81

When embedded in genes, USS favors low conservation areas



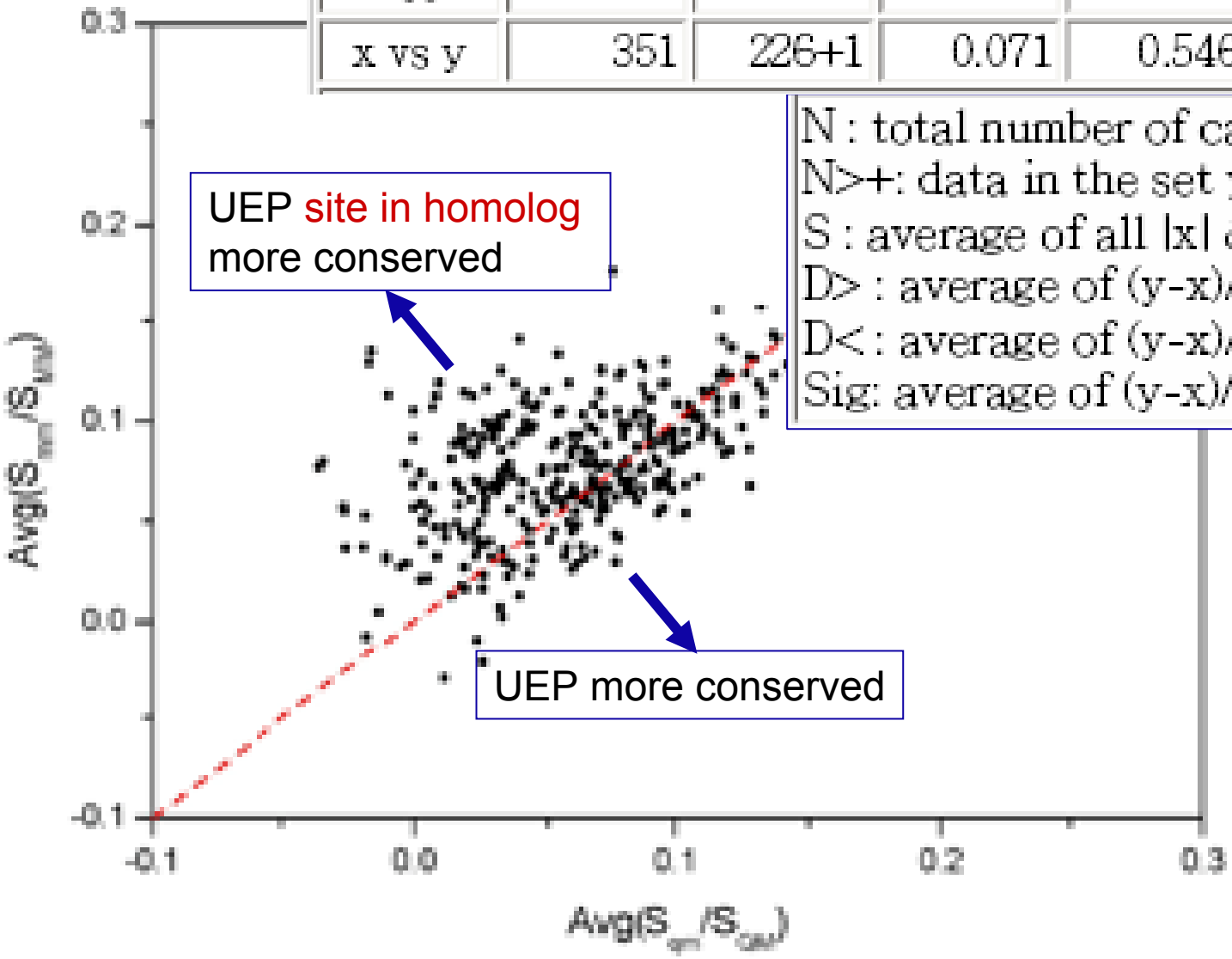
Homolog search of UEP embedded gene by BLAST



Embedding of USS reduces conservation of UEP site

Test	N	N>+	S	D>	D<	Sig
x vs y	351	226+1	0.071	0.546	-0.248	0.265

N : total number of cases in figure
N>+: data in the set y >= x
S : average of all |x| & |y|
D> : average of (y-x)/S over set y >= x
D< : average of (y-x)/S over set y <= x
Sig: average of (y-x)/S over all data



Two views on "How did USS emerge?"

- USS first:
  - Naturally competent bacteria had a preference to bind to USS; high USS content is a result of recombination of uptaken DNA fragments containing USS
  - This begs the question: how did the "preference to bind" emerge?
- Preference first:
  - Conspecific (homologous to self) DNA is more beneficial than nonconspecific DNA; the USS evolved as a signal to allow bacteria to tell one from the other

Emergence of USS

Central assumption: Uptake of conspecific DNA is more beneficial than uptake of other DNA.

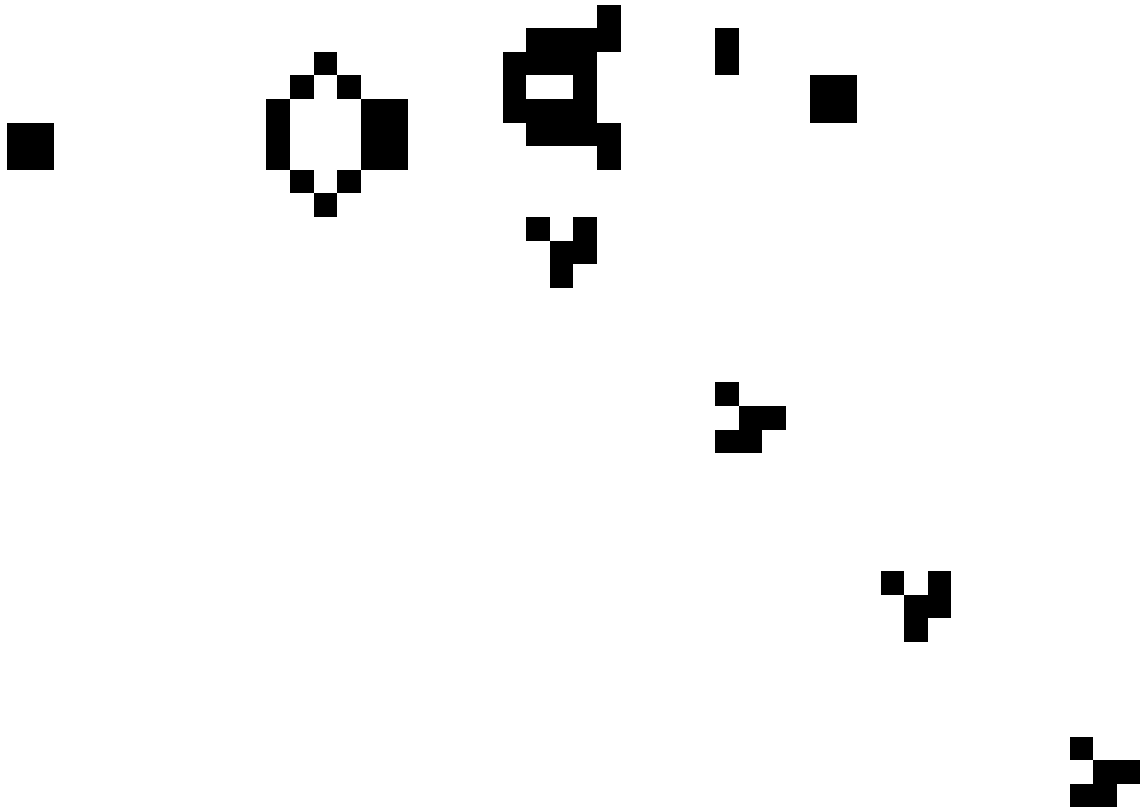
Can we demonstrate the emergence of USS in a computational model?

Agent-Based Model (Cellular automata)

- Reality is complex, but models need not be
- Von Neumann machines - a machine capable of reproduction; the basis of life is information
  - Stanislaw Ulam: build the machine on paper, as a collection of cells on a lattice
  - Von Neumann: first *cellular automata*
- Conway: Game of Life
- Wolfram: simple rules can lead to complex systems

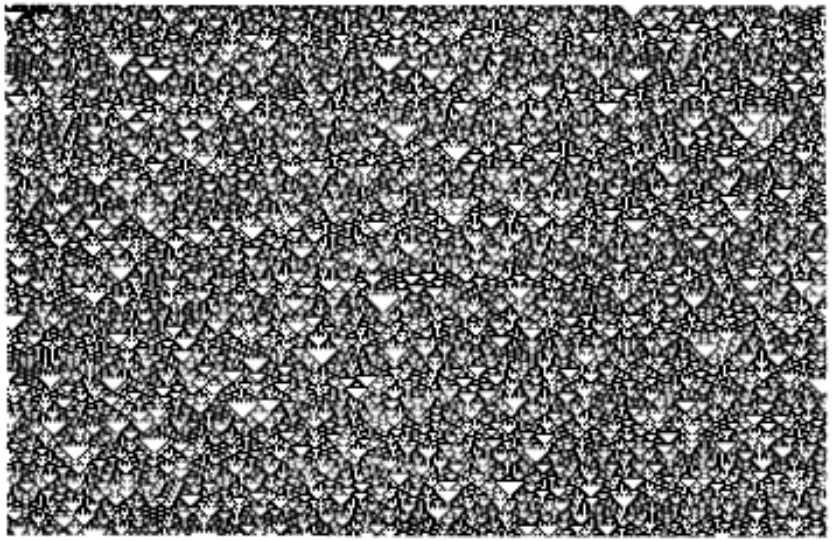
Conway's Game of Life: Simple rules may generate not-so-simple patterns

Gospers glider gun

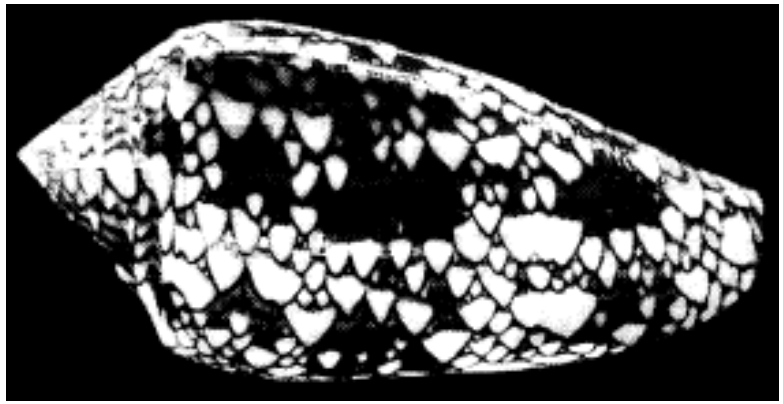


Wolfram: Self-Organization in Cellular Automata

Cellular Automata



Mollusk shell



An Agent-Based Model for emergence of USS

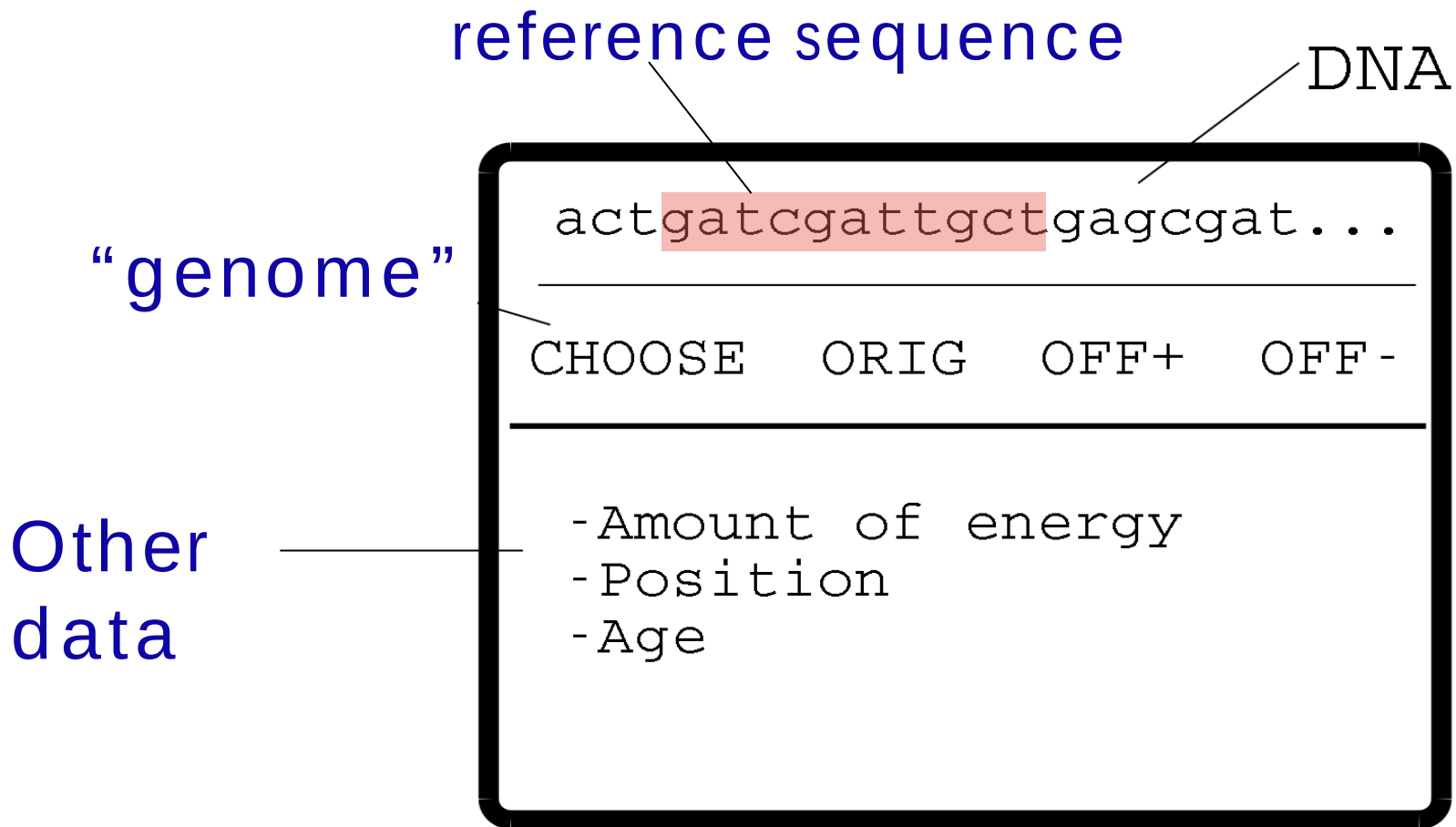
- Uptake of conspecific DNA beneficial
- Uptake of alien DNA not detrimental
- Alien DNA is random
- Initial conspecific DNA is random as well
- Agents must learn to distinguish between conspecific and alien DNA

Structure of the Model

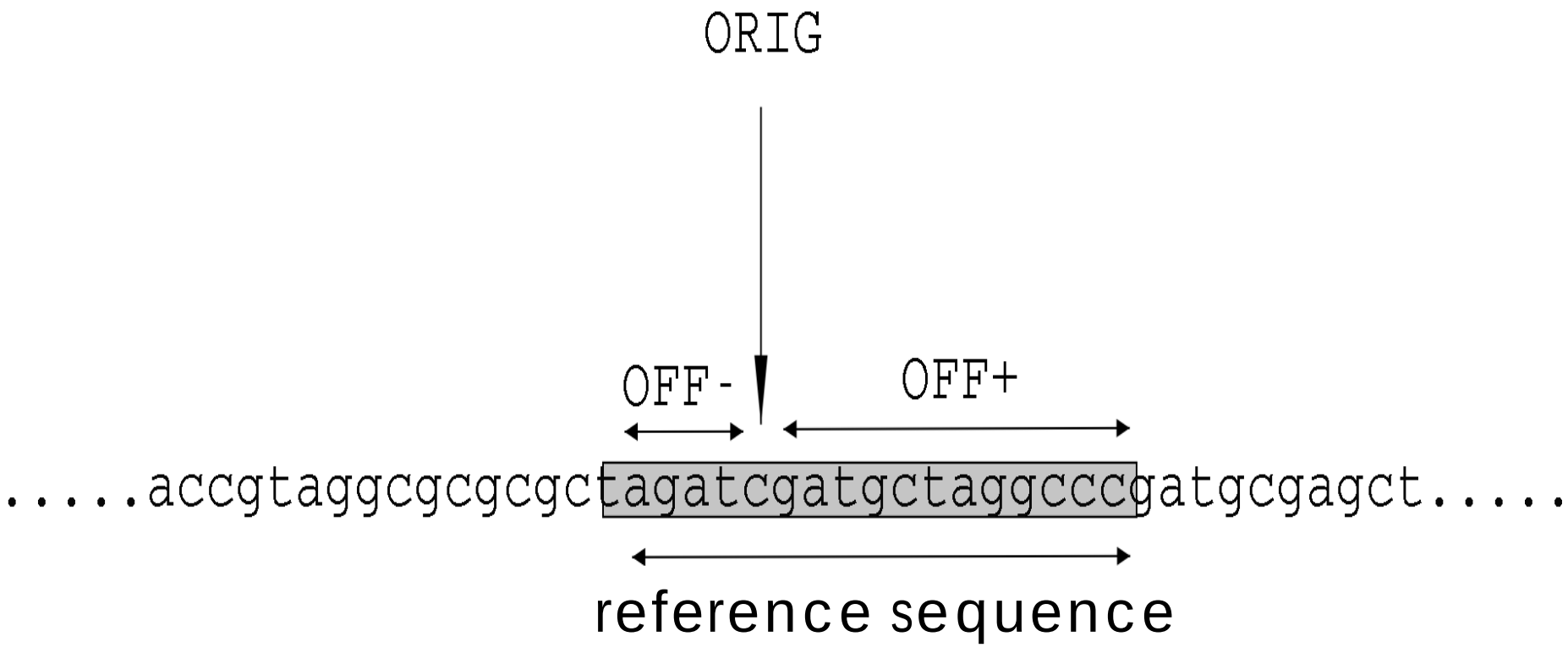
- Agents (the genomes)
  - Compare: agents in Conway's and Wolfram's games are just regular geometric shapes (squares or triangles)
- Environment
- Rules

AGENTS

An agent; a DNA sequence; a “genome” with a reference sequence; house-keeping data



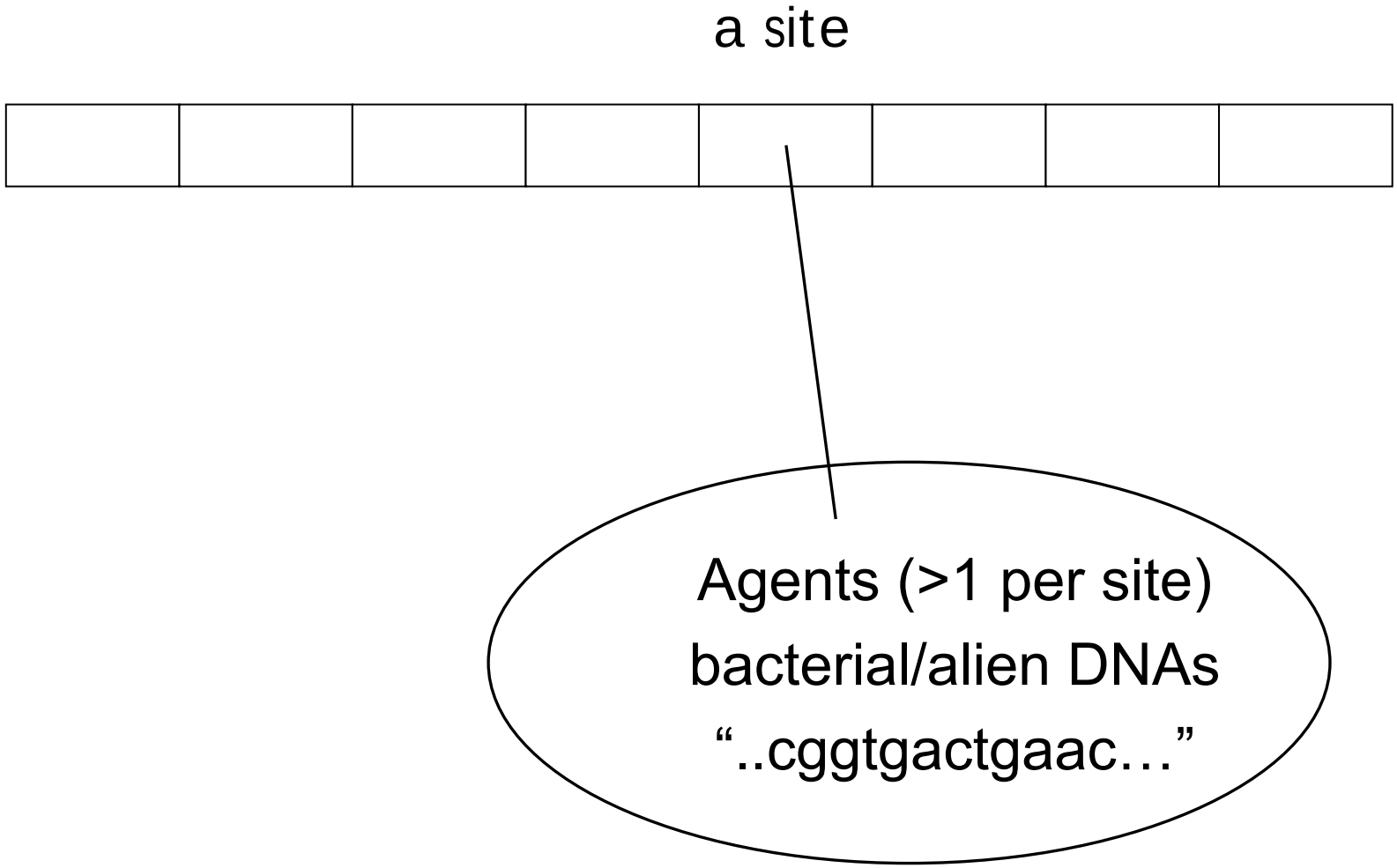
“Genome” contains information pointing to the reference sequence



ENVIRONMENT

- 1-dimensional lattice with N sites
- Each site may have more than one agent
- A site also contains two types of DNA fragments
  - "Bacterial" DNA (from dead agents)
  - "Alien" DNA (continuously replenished)
  - All fragments have same fixed length

Each sites may have more then one agent



Two types of DNA fragments

RULES

- Time
  - system progress by discrete time steps
  - In each time step updating loops through all agents
- Updating operations at each step
  - Feeding
  - Reproduction
  - Death
  - Mutation
  - Refill alien food

Feeding rules

- Each agent presented with fixed number of fragments
  - Always enough alien fragments available
  - If available, bacterial fragment presented with low probability.
    - ►NB! Bacterial fragments will often be taken from ancestor of agent!
- Each agent takes exactly 1 fragment/time-step
  - If CHOOSE == false, then accept first item.
  - Else, compare fragments with reference (one by one).
    - ►Accept food if reference sequence is contained in fragment or last fragment encountered.
- Once food is accepted, agent aborts inspection of further fragments
- Food is converted to “energy” after uptake
  - Bacterial DNA has higher energy than alien DNA

Reproduction rules

- Agent reproduces if energy exceeds preset threshold
  - ½ energy given to offspring
  - Offspring is placed in same or neighboring site
- If maximum population size reached, then for every new born agent, an old one must die

Mutation rules

- DNA and Genome of agent at birth may be mutated
  - DNA – one of following two
    - Point mutation: randomly change a letter at a randomly selected site
    - Copy mutation: replace a randomly selected target substring by a randomly selected source substring of the same length
  - Genome – change either size or location of the reference sequence by one unit

Death rules

- Agent dies if...
  - it runs out of energy (never happens)
  - lives beyond a preset age
  - killed because it is the oldest when maximum population size reached and new agent is born

Simulations

- Initially DNA of agents is random
- Agents cannot distinguish between bacterial and alien fragments
- Get fit: Eat your ancestors!
  - Have (short) repeated subsequences on DNA.
  - Set reference sequence to one of those repetitions.
- Get fitter!
  - Because of limited space, agents must keep evolving ("Red Queen Effect").

Run parameters

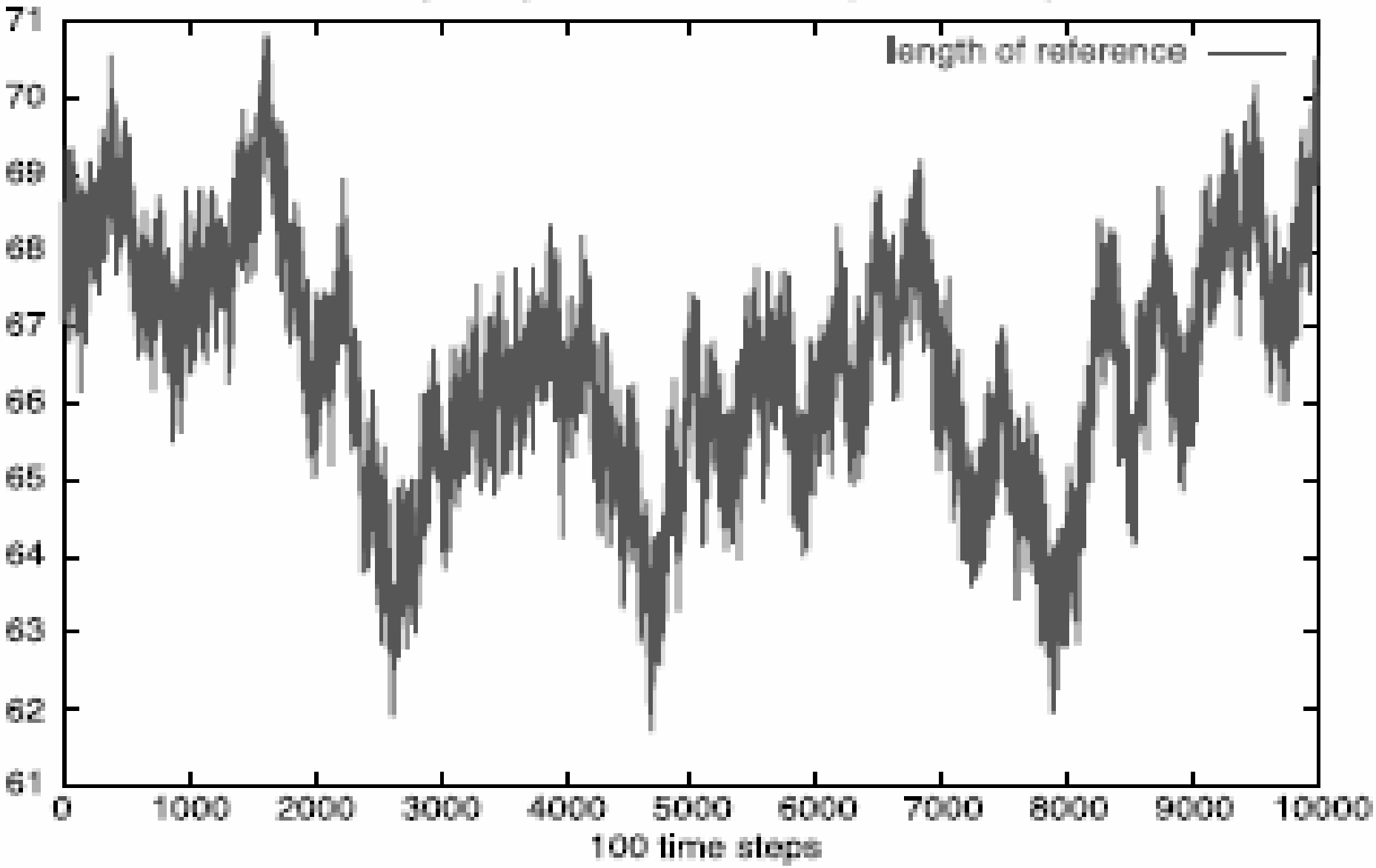
DNA length	10,000
World size	30
Max. population size	300
Mutation rate	0.9
Point-mutation rate	0
Maximum length of copied substring	**300**
Max. no. of fragments presented to agents	20
Size of fragments	100
Min. energy to reproduce	6
Max. lifetime	10
Payoff for alien fragments	1
Max. no. of bacteria per site	200

Run 1

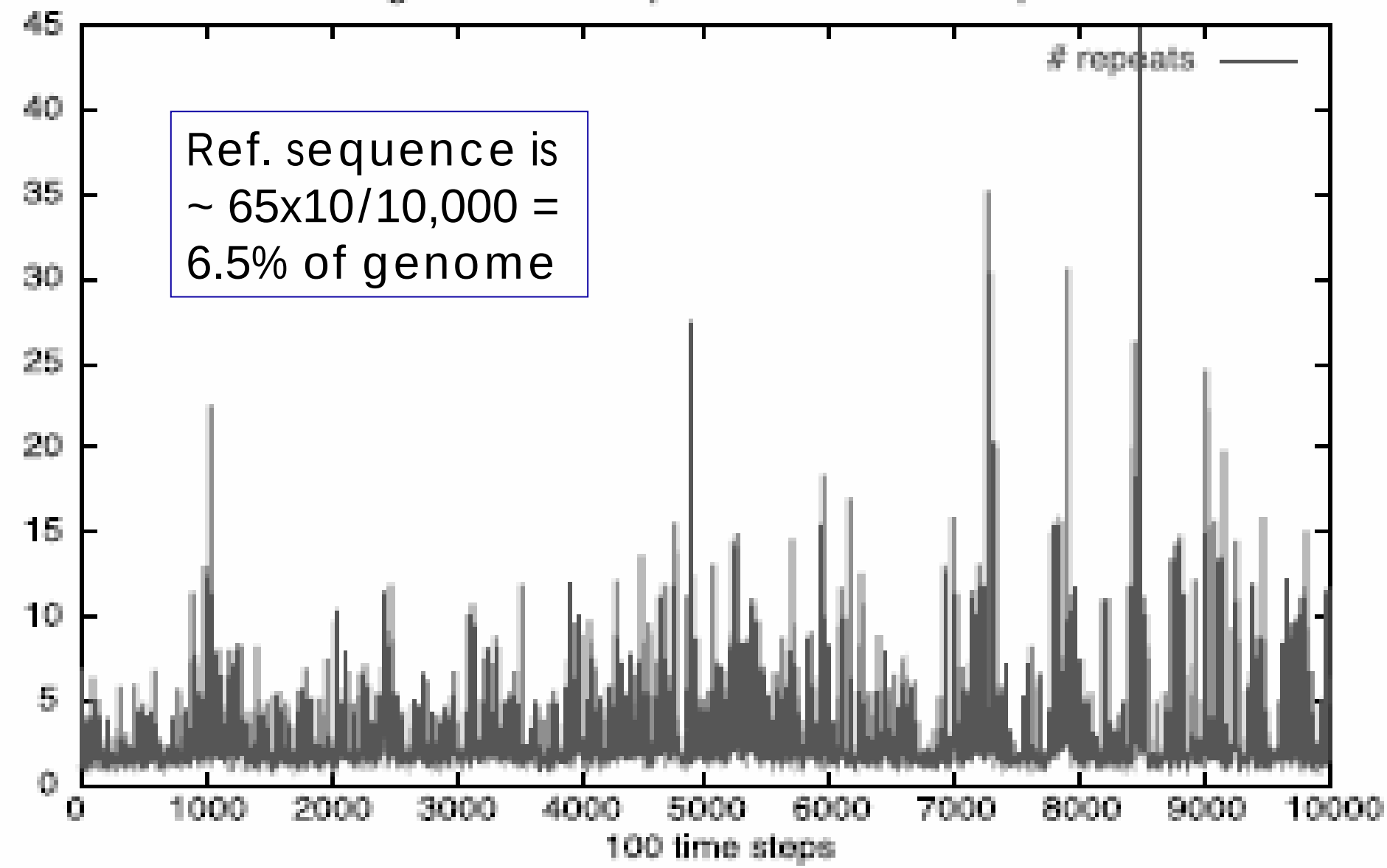
Number of updates = 1,000,000
Energy from alien fragment = 1
Energy from bacterial fragment = 2

Nothing happens

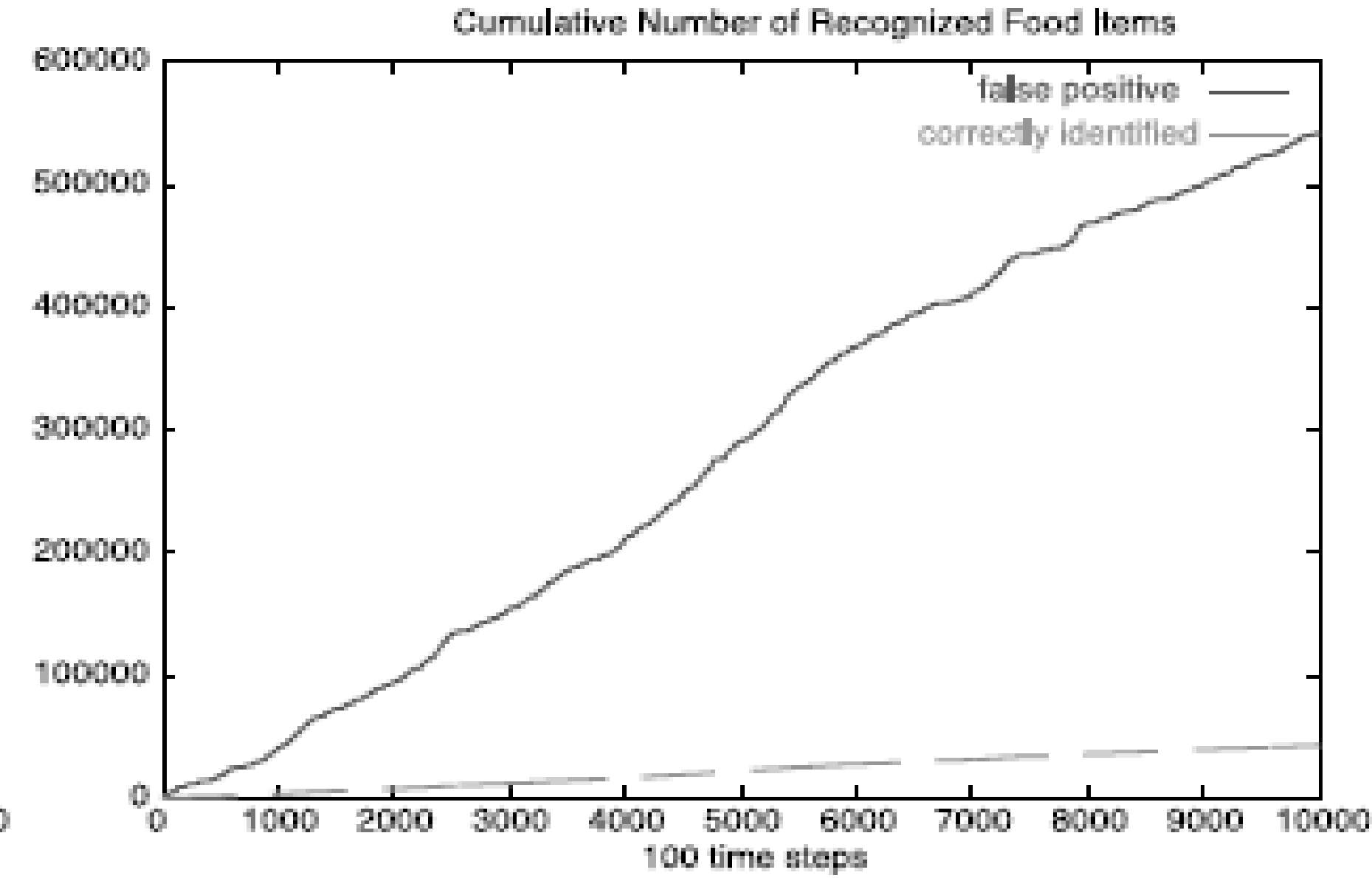
Average length of reference sequences



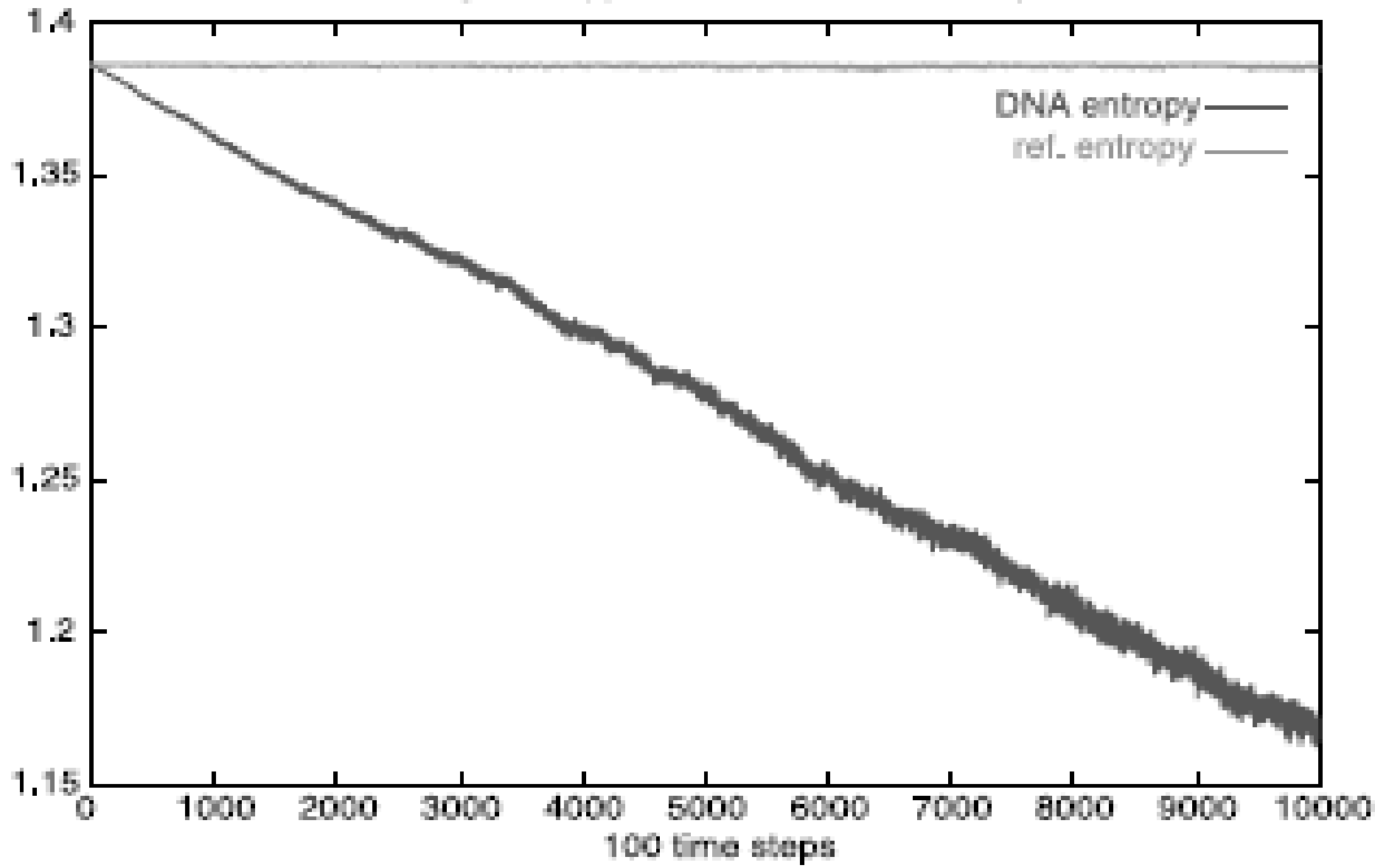
Average number of repeats of reference sequences on DNA



Cumulative number of recognized uptakes



Average entropy of DNA/reference in population

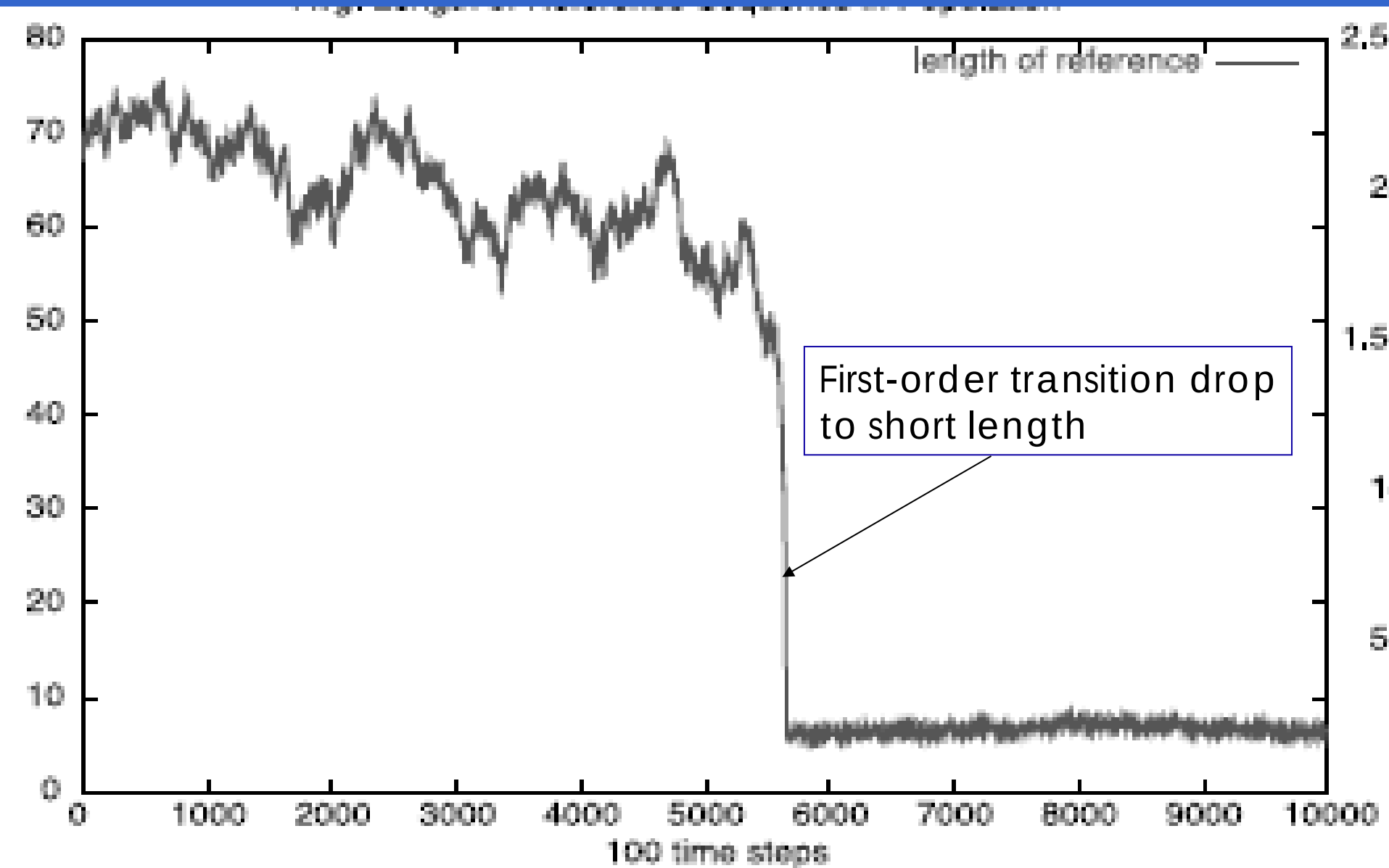


Run 2

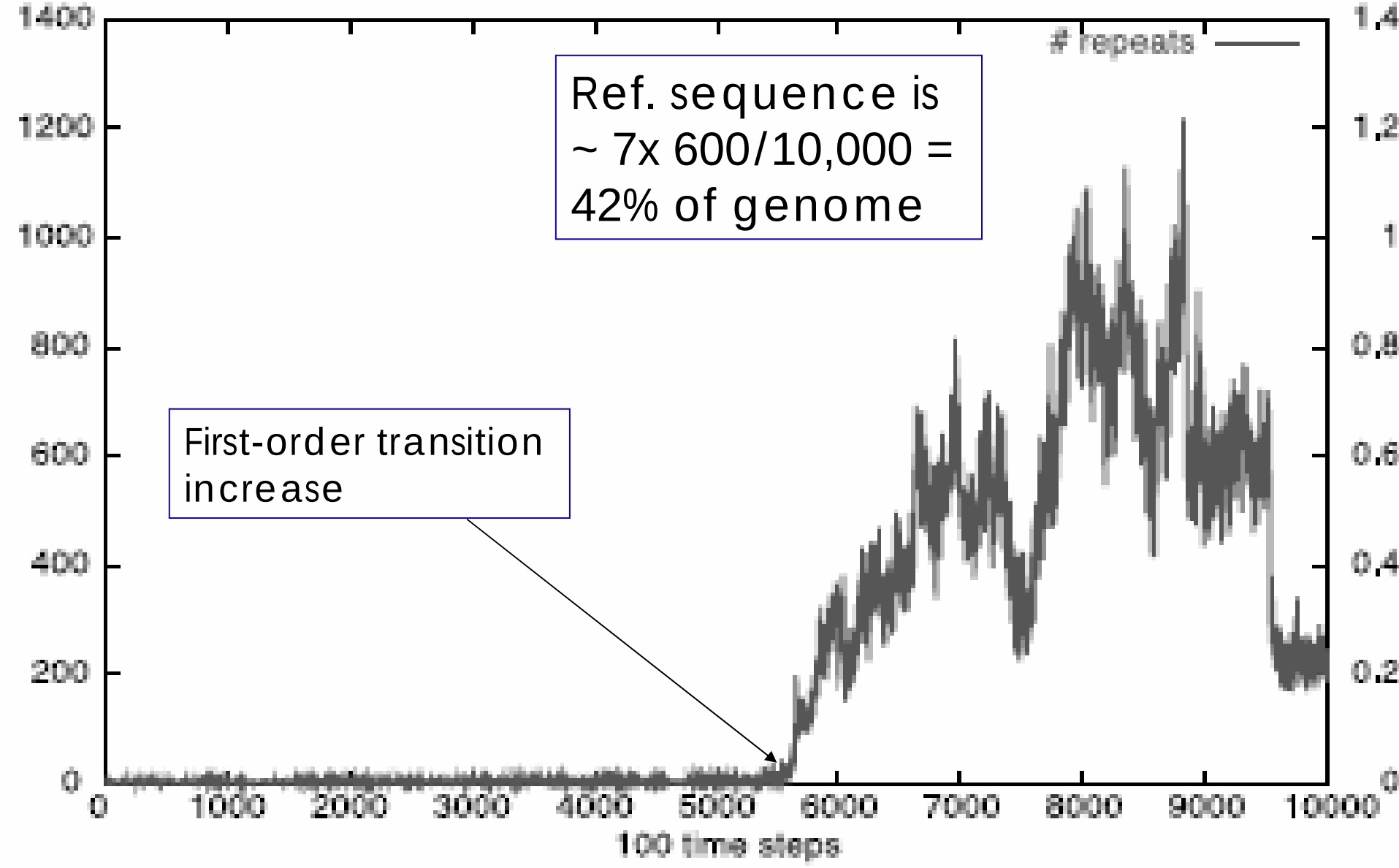
Number of updates = 1,000,000
Energy from alien fragment = 1
Energy from bacterial fragment = 3

Spontaneous emergence of USS

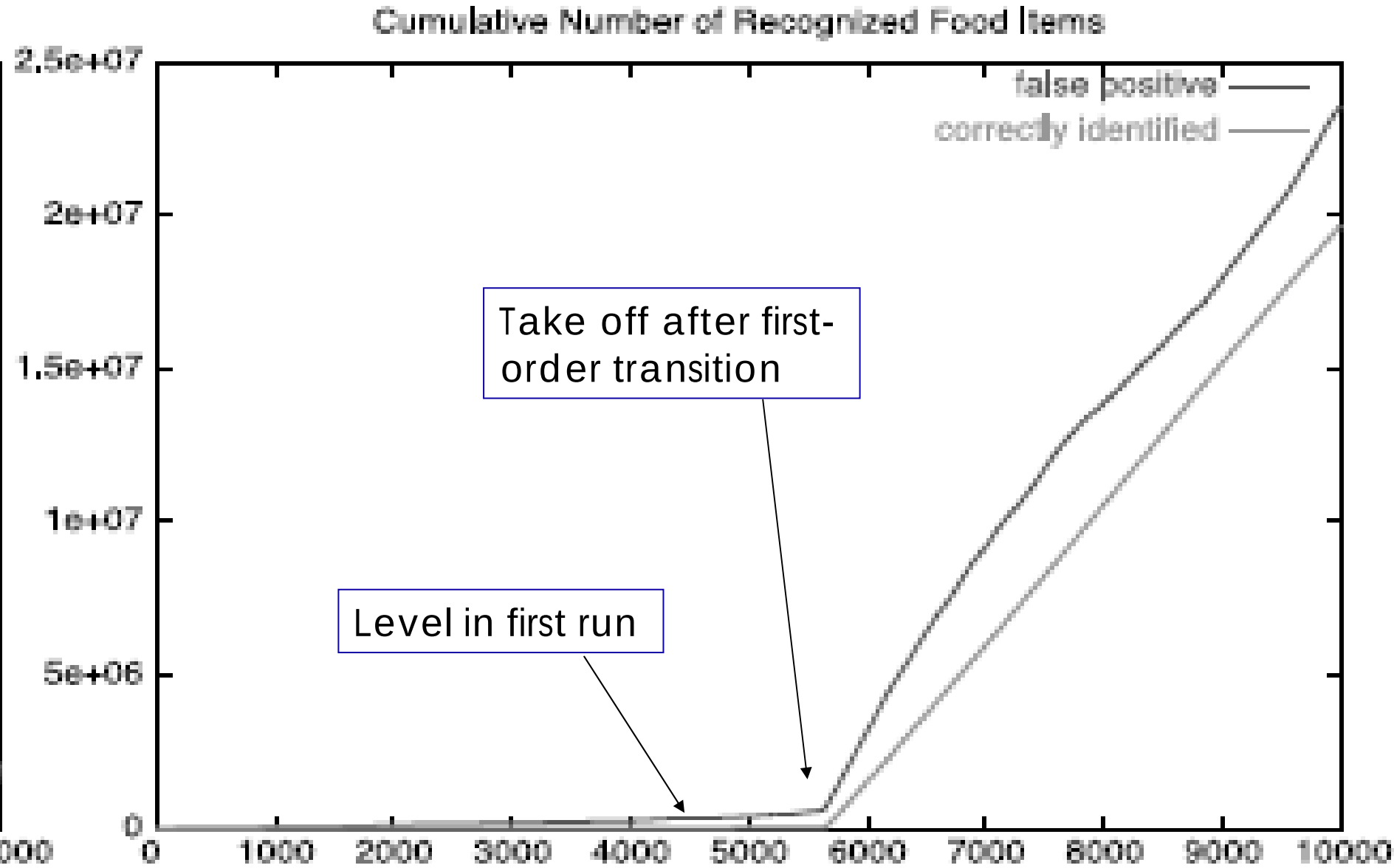
Average length of reference sequences



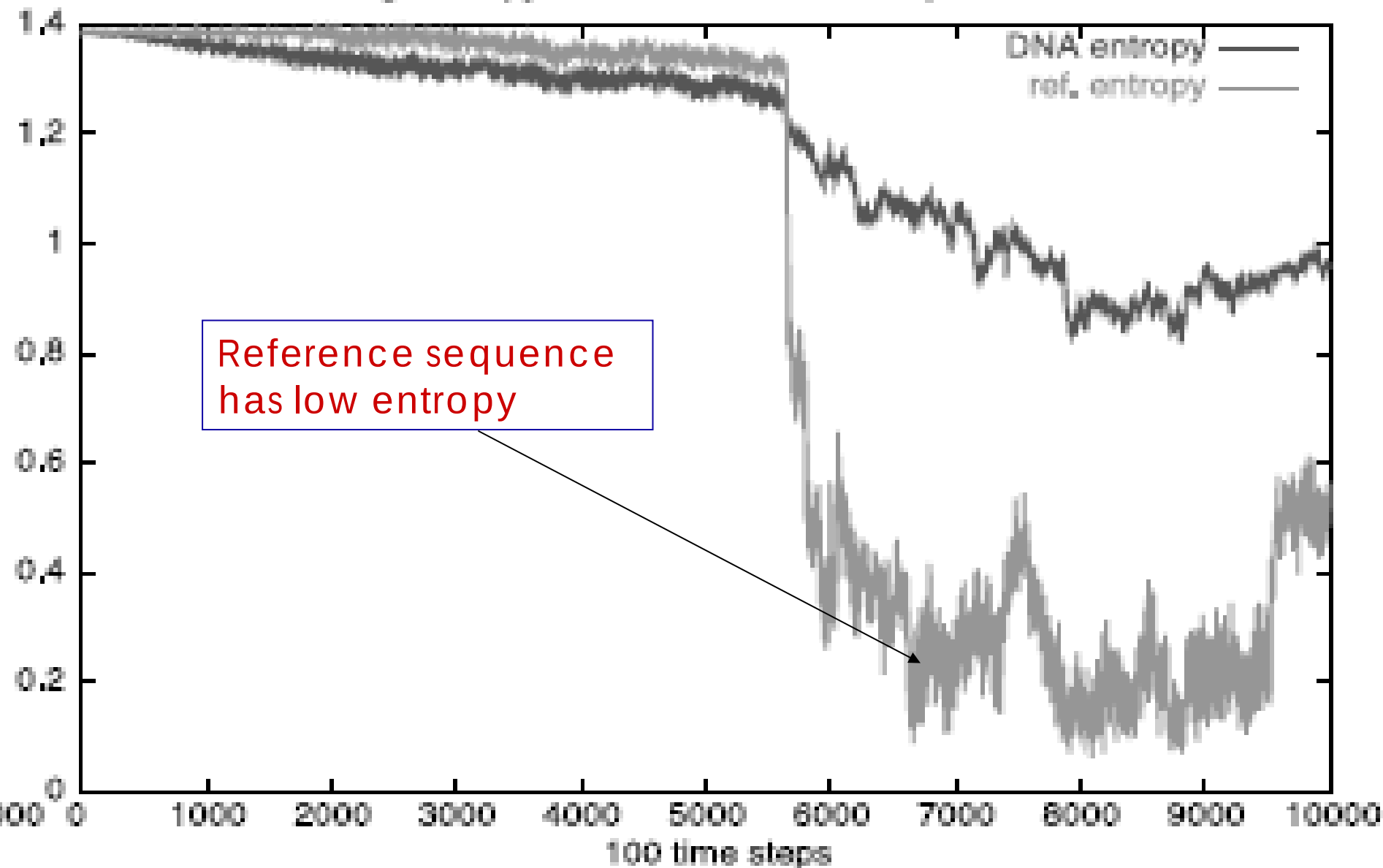
Average number of repeats of reference sequences on DNA



Cumulative number of recognized uptakes



Average entropy of DNA/reference in population



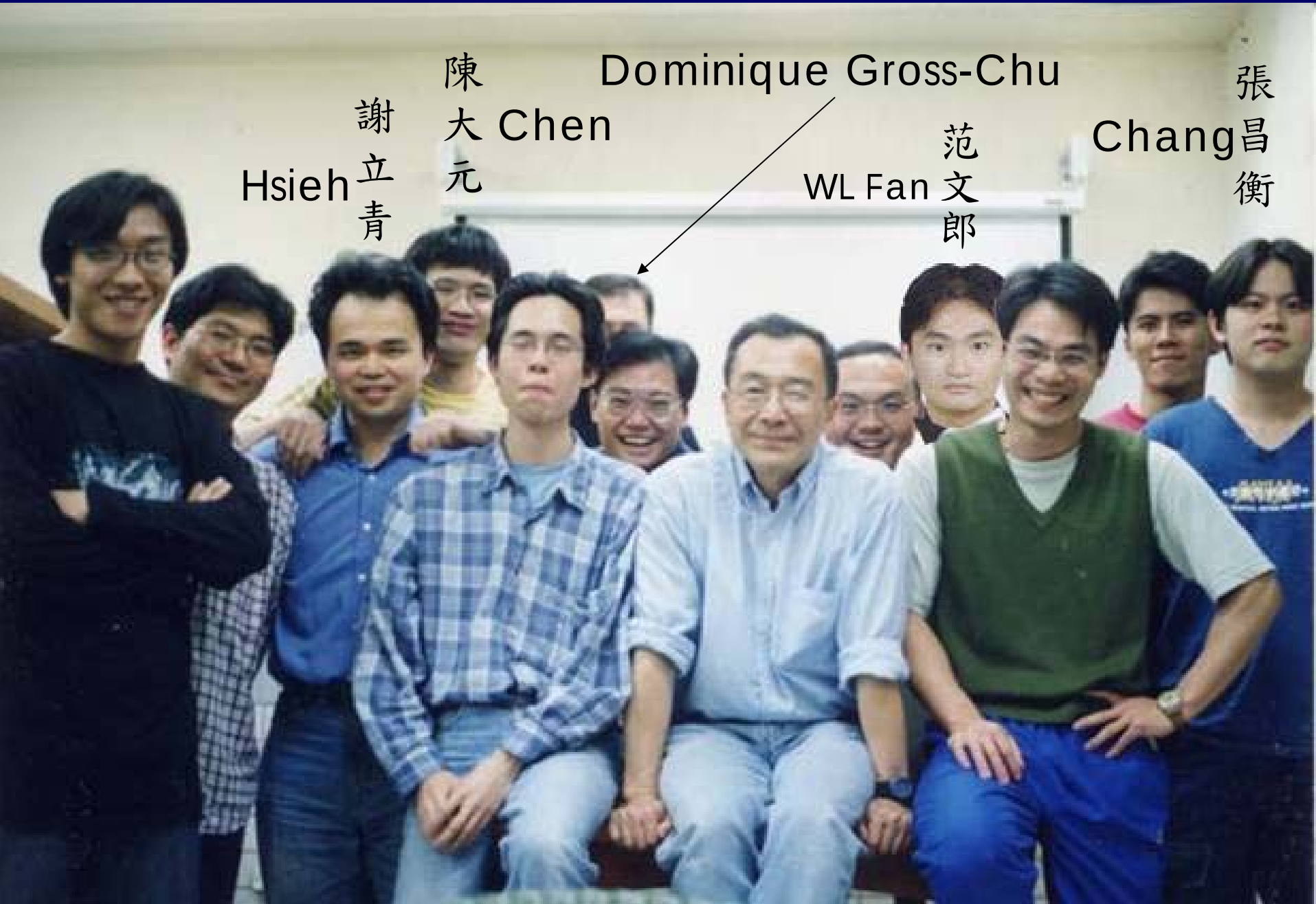
Summary

- In the simple model uptake signal sequences (USS) can spontaneously emerge provided conspecific fragments are moderately more favored than alien fragments
- Uniformity symmetry was spontaneously broken
- The emergence of USS is a first-order transition in time (inflation)
- Prototype of emergence in biology?
- Analytic model?

People

- Genome growth
  - **Li-Ching Hsieh** (now at Genome Center, AS) and other members of CBL
- USS - Biology
  - **Da-Yuan Chen**, He-Hsin Cancer Research Hospital, Taipei
  - **Rosey Redfield**, Zoology, U. British Columbia, Canada
  - **Mohamed Bakkali**, Genetics, U. Nottingham, UK
- USS - Cellualar automata
  - **Dominique Chu/Gross**, Comp. Sci., U. Kent, UK
  - **Tom Lenaerts**, U. Libre de Bruxelles, Brussels, Belgium

Computational Biology Laboratory (2003)



Our papers are found at Google: HC Lee

Thank you!