

The Blind Self-Copier –
Universality, inflation & spontaneous symmetry
breaking in the growth of genomes

National Tsing-Hua University

Shinchu, Taiwan

2006 March 8

HC Lee

Computational Biology Lab

Dept. Physics, Inst. Biophysics & Inst. Systems Biology

National Central University

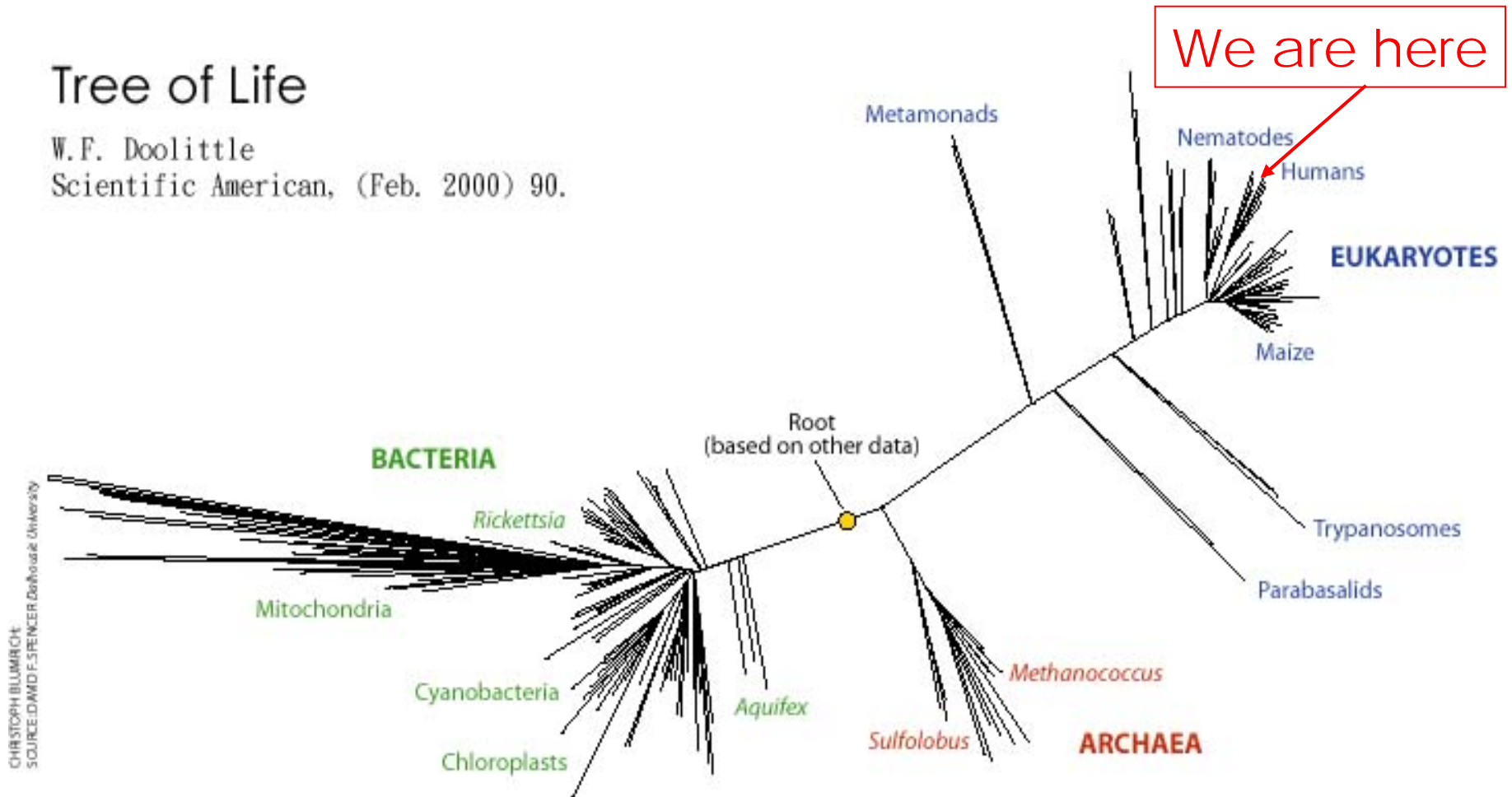
Some concepts to be discussed

- Randomness and order
- Second law of thermodynamics in genome growth
- Diversity and universality
- Genome grew by random self-copying
- Inflation in early genome growth?
- Genome evolution by Cellular automata
- Emergence by spontaneous symmetry breaking in evolution

Life is highly diverse and complex

Tree of Life

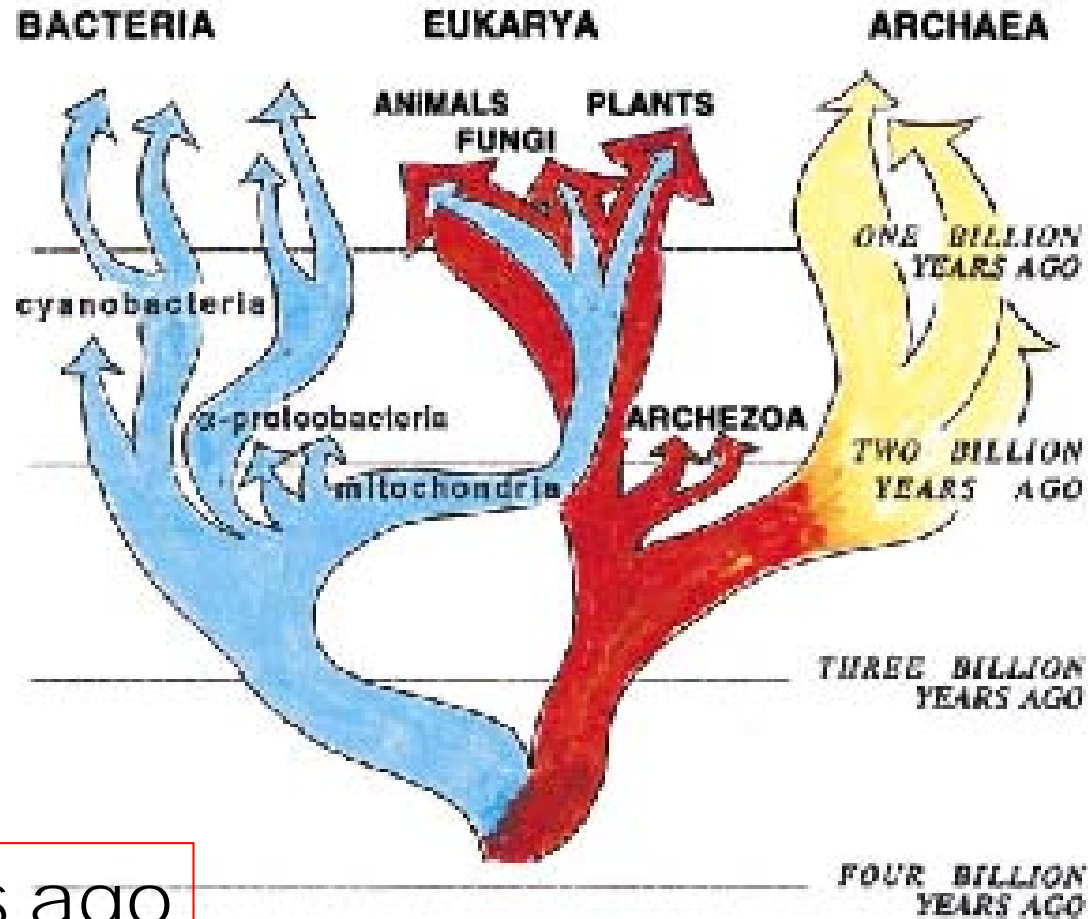
W.F. Doolittle
Scientific American, (Feb. 2000) 90.



And it took a long time to get here

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

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 chromo-
some of
Homo sapien

[http://
www.ncbi.nlm.nih.gov/
entrez/viewer.fcgi?val
=2276452&db
=Nucleotide
&dopt
=GenBank](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 tgtttctct tcccaaaag acacttegca gccctcttg
61  ggatccagcg cagcgcaagg taagccagat gcctctgctg ttgccctccc tgtgggctg
121 ctctctcac gccggcccc acctgggcca cctgtggcac ctgccaggag gctgagctgc
181 aaaceccaat gaggggcagg tgctccgga gacctgctc ccacacgcc atcgttctgc
241 cccggcttt gaggctctc aggccctct gtgcacctc ccctagcagg aacatgccgt
301 ctgccccct gagctttgca aggtctcgg gataatagga aggtctttgc cttgcaggga
361 gaatgagtea tccgtgctc ctccgagggg gattctggag tccacagtaa ttgcagggct
421 gacactctgc cctgcaccg gcgccccag tctccccac ctctctctc catcctgtc
481 tccggctatt aagacggggc gctcaggggc ctgtaactgg ggaaggtata cccgccctgc
541 agaggtggac cctgtctgtt ttgattctg tccatgtcc aaggcaggac atgacctgt
601 tttggaatgc tgattatgg atttccagg cactgtgcc ccagatacaa tttctctga
661 cattaagaat acgtagagaa ctaaatgeat tttctctta aaaaaaaaaa aaacaaaaa
721 aaaaaaaaaa aaacaaaaa actgtactta ataagatcca tgcctataag acaaaggaac
781 acctctgtc atatatgtg gacctcgggc agcgtgtgaa agttacttg cagtttgcag
841 taaaatgaca aagctaacac ctggcgtgga caatcttacc tagctatgct ctccaaaatg
901 tatttttct aatctgggca acaatgggtc catctcgggt cactgcaacc tccgctccc
961 aggttcaagc gattctccg cctcagcctc ccaagtagct gggaggacag gcaccgccca
1021 tgatgcccgg ttaattttg tatttttagc agagatgggt ttcgccatg ttggccaggc
1081 tggctcga a ctctgacct caggtgatcc gcctgcctg gcctcccaaa gtgctgggat
1141 gacagcgctg agccaccgag ccagccagg aatctatgca ttgccttg aatattagcc
1201 tccactgcc catcagcaa aggcaaaaca ggtaccagc ctccgccac ccctgaagaa
1261 taattgtgaa aaaatgtgga attagcaaca tgttggcagg attttctg aggtataag
1321 ccacttctt catctgggtc tgagctttt tgtattcgg ctaccattc gttggttctg
1381 tagttcatgt ttcaaaaatg cagcctcaga gactgcaagc cgtgagtc aatacaata
1441 gatttttaa gtgtattat ttaaacaaa aaataaaatc acacataaga taaacaaaa
1501 cgaaactgac ttatacagt aaataaacg 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 \vec{I}

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

*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

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

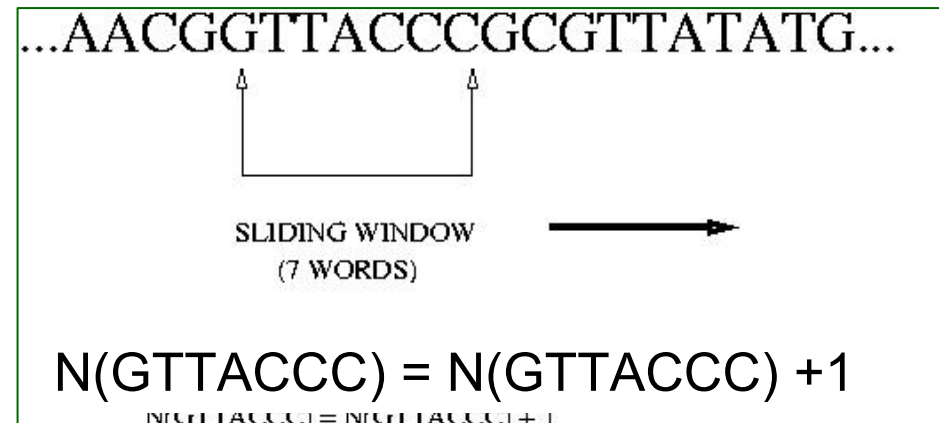
def $(CV_{random})^2 = \tau/L$

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

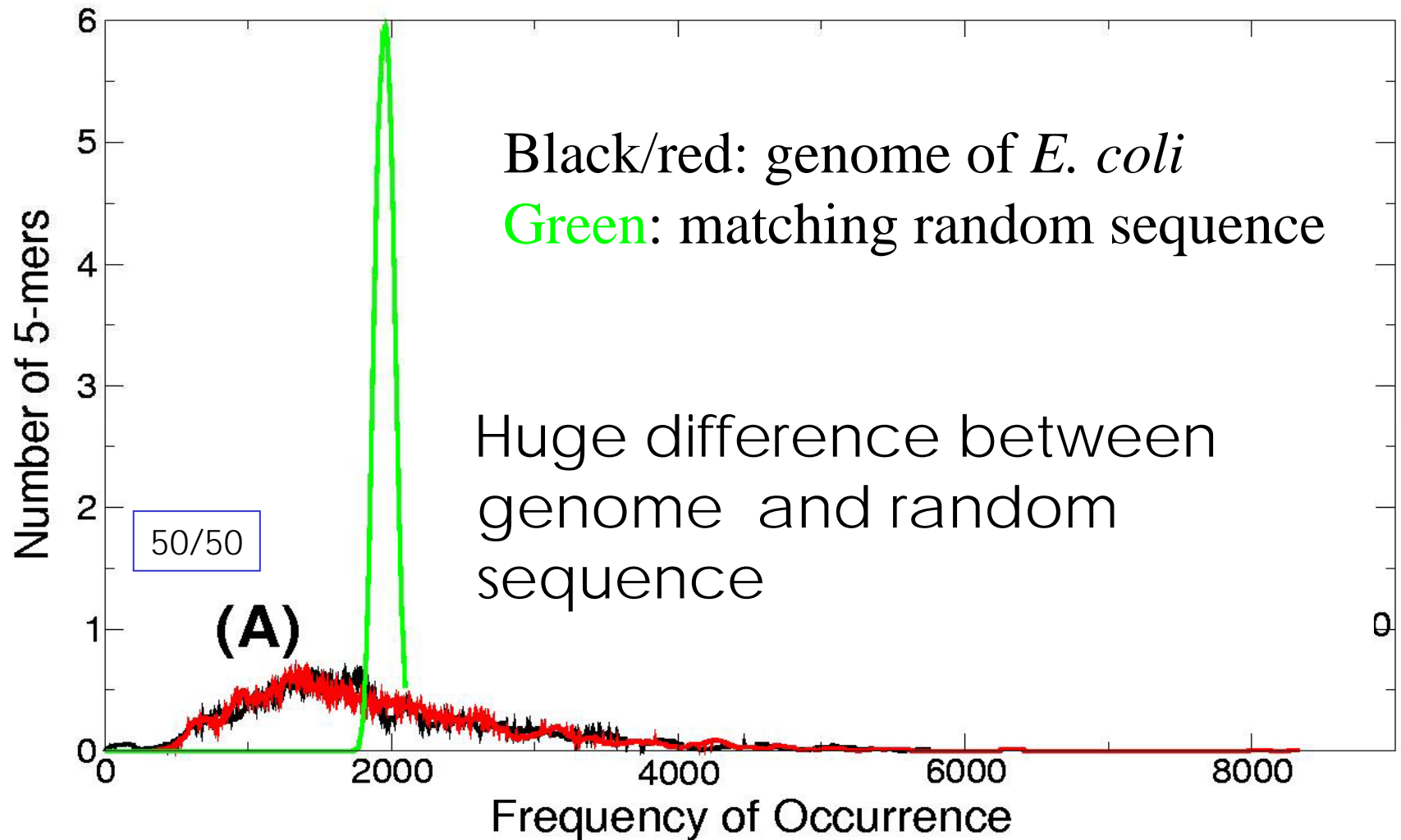
Note: technical details when p not equal to 0.5

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

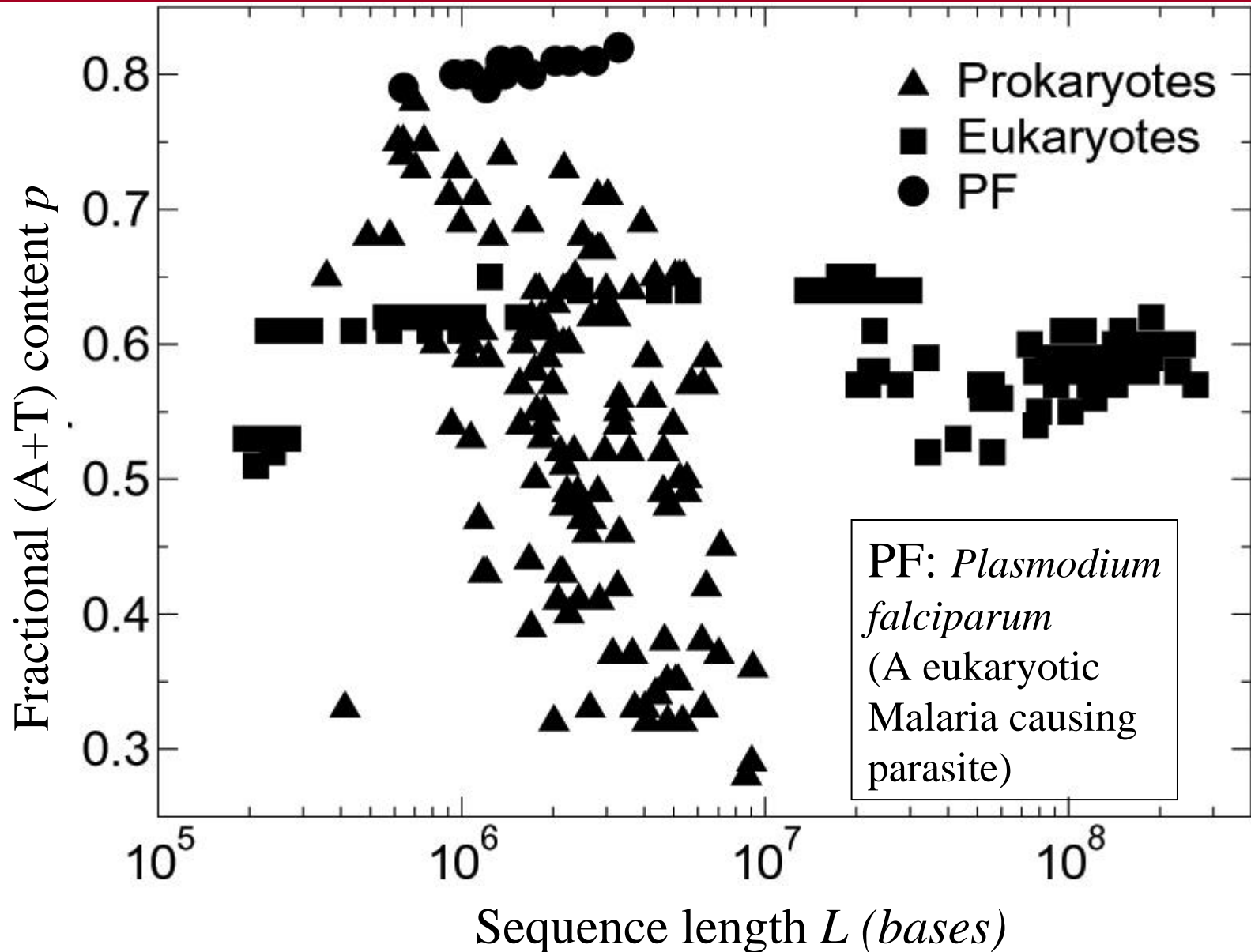
$$L_{eff} = \tau / (CV)^2$$

- 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 grows as L

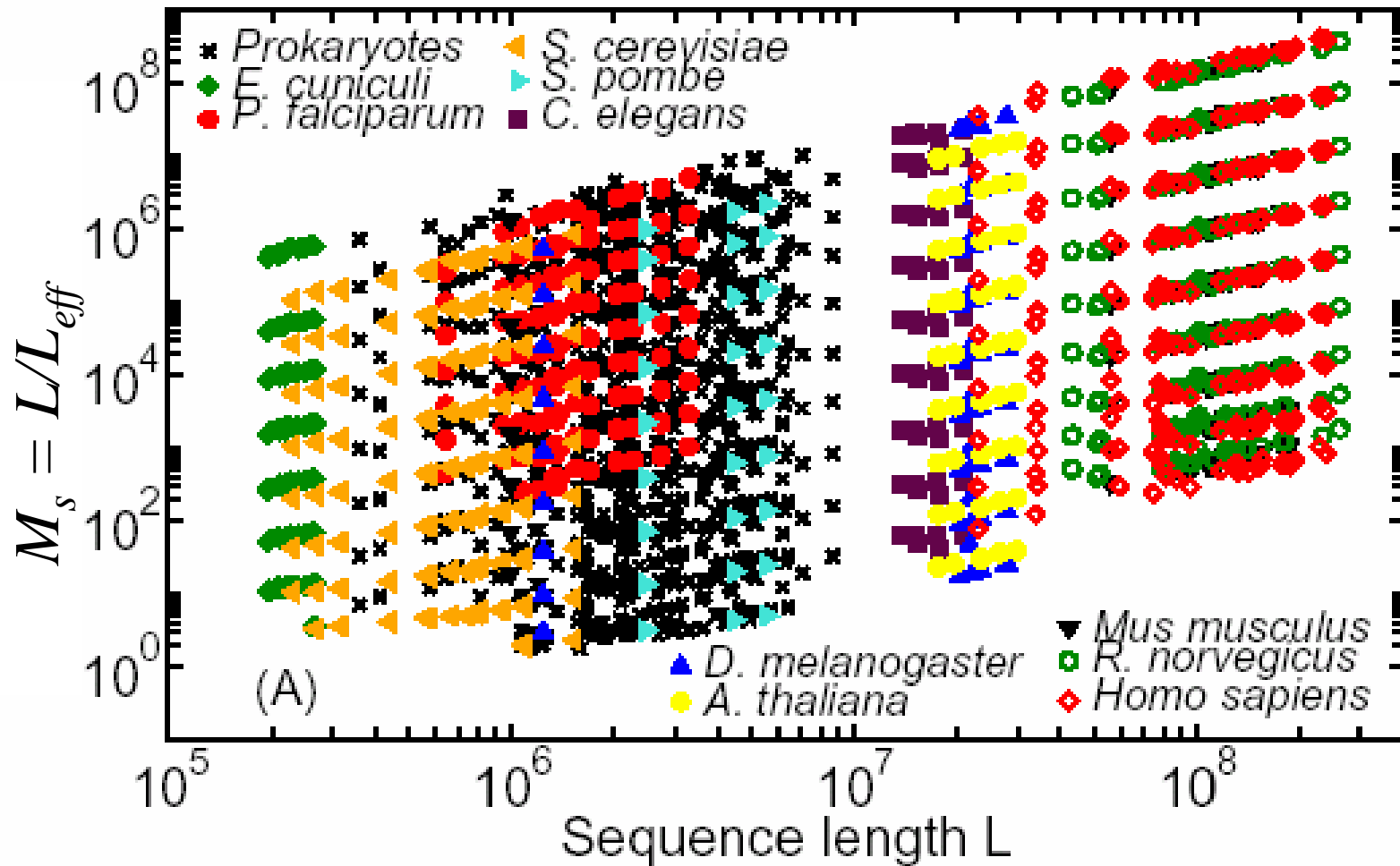
The genome data set includes all complete sequences

- 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 $L_{\text{eff}}(k)$ for $k= 2$ to 10 (or to k_{max} such that sequence length $> 4^{k_{\text{max}}}$)
 - Each k has about 278 pieces of data
 - All told about 2000 pieces of data

Complete Genomes are diverse

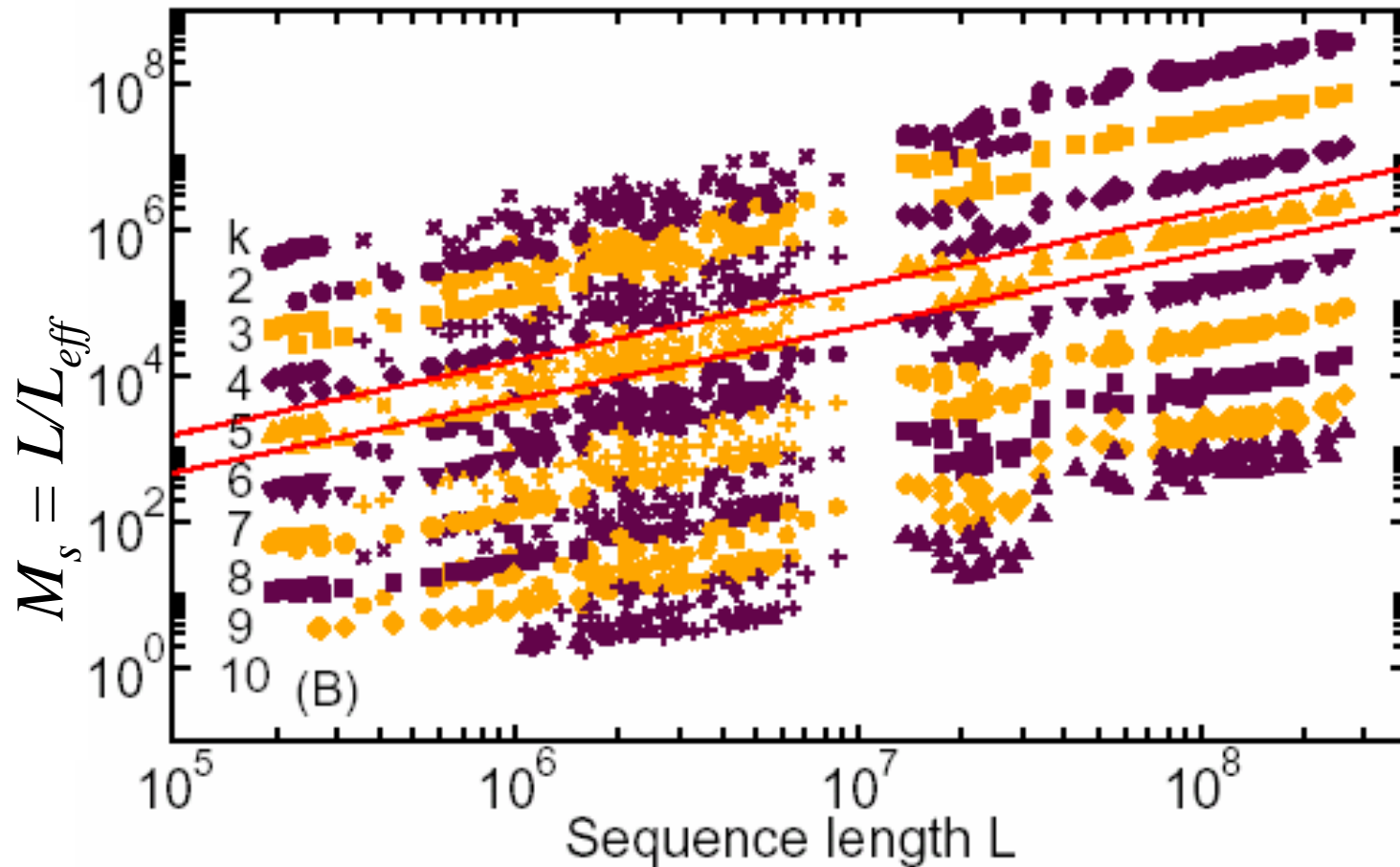


Results: color coded by organisms



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 $M_s \sim L$ " k band".

Genomes are in Universality Classes

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

- Obeys

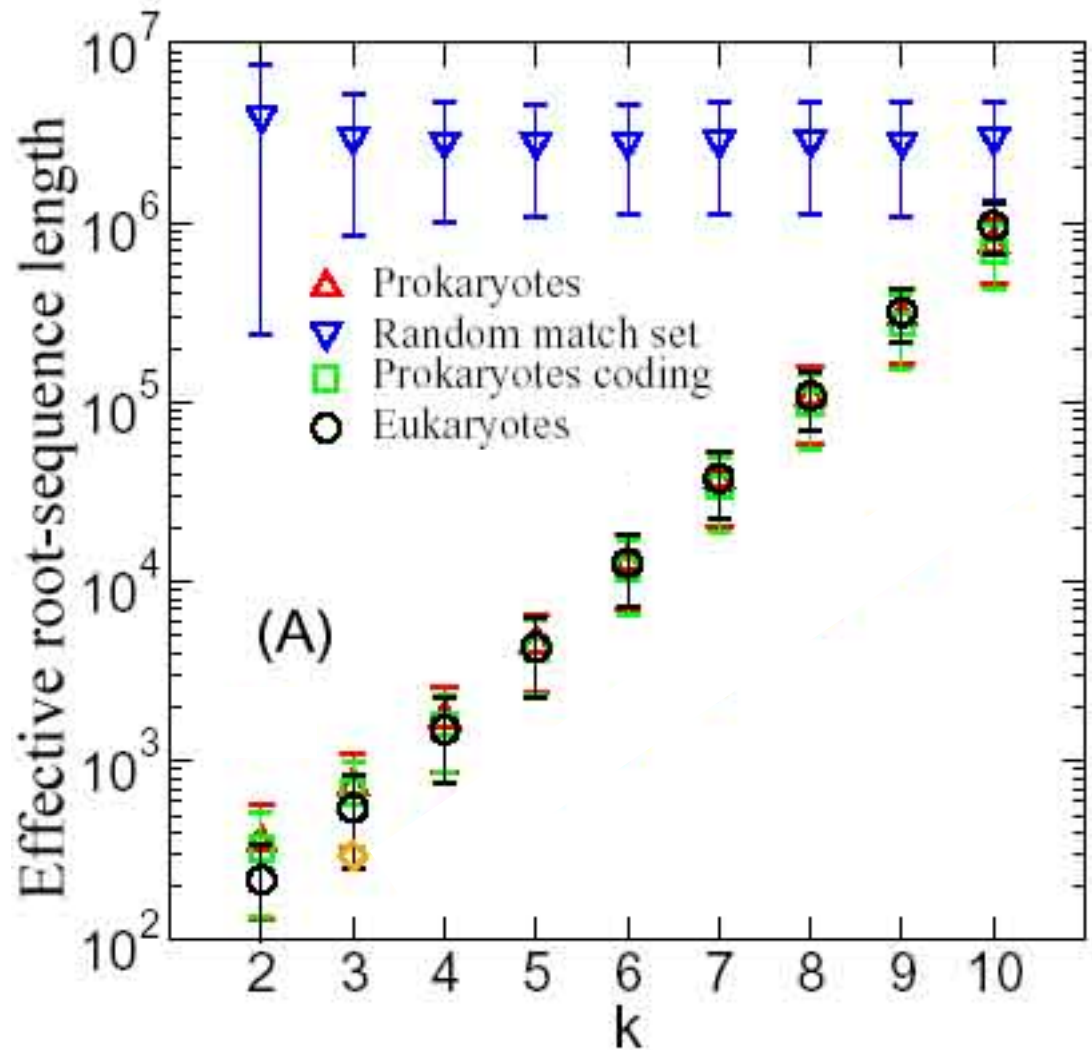
$$\log L_r(k) = a k + B$$

1989 pieces of data given by two parameters.

$$a = 0.398 \pm 0.038$$

$$B = 1.61 \pm 0.11$$

- 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
 - $\text{Log } L_{eff}(k) = ak + B$
 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

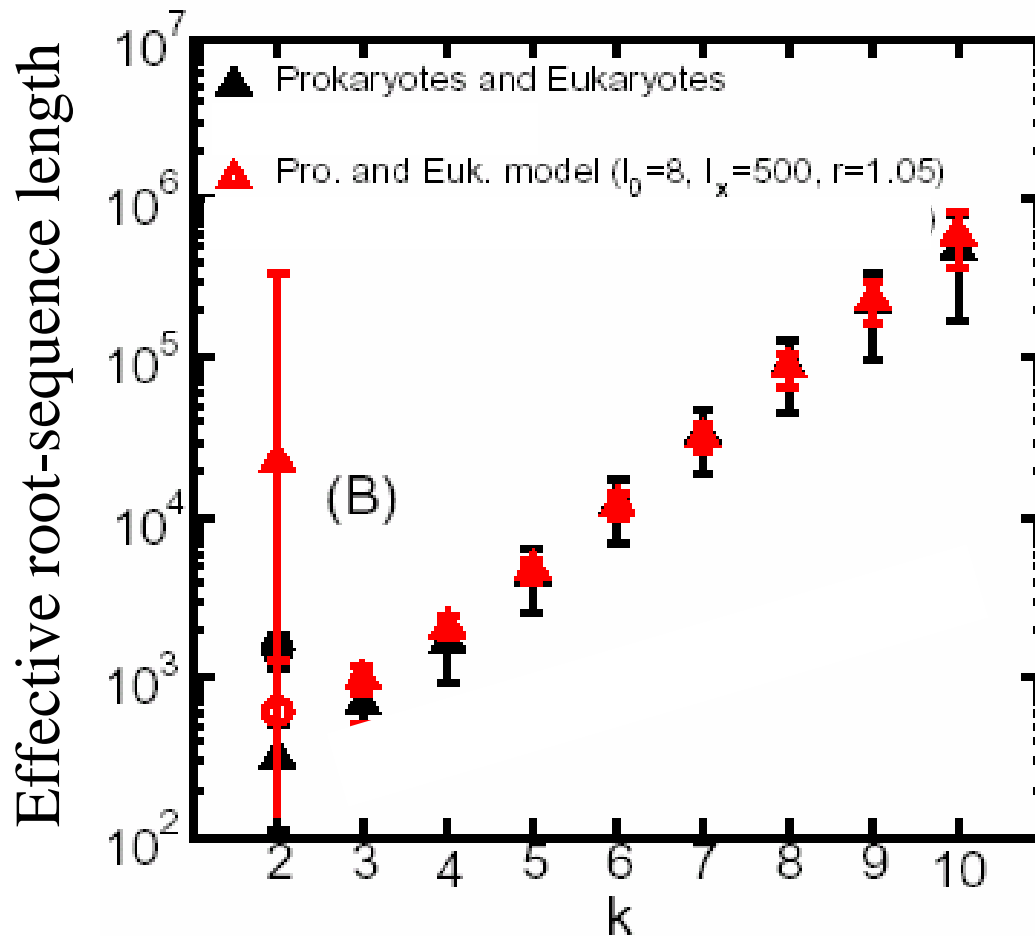
A Universal Model for Genome Growth

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)

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

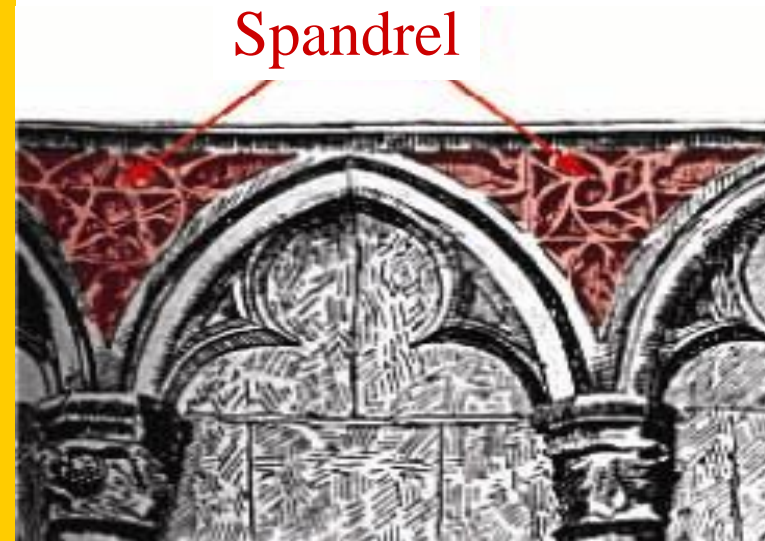
Many biological phenomena explained by model

- Preponderance of **homologous genes** in all genomes
- Genome is full of **non-coding repeats**
- Large-scale genome “**rearrangements**”
- Rapid **rate of evolution** - random self-copying is an extremely efficient way for information accumulation; it is genome’s way to “beat” the **2nd law of thermodynamics**
- 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

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 *Tetrahymena* (纖毛蟲) contains ~50 copies of each gene, except for rRNA genes, which exist in ~10,000 copies.

Evolution of genome –
emergence &
spontaneous symmetry
breaking

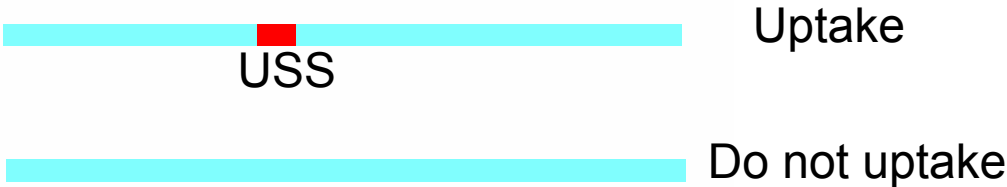
DNA uptake signal sequence

- 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 $\sim 1/\text{kb}$, expected frequency is $\sim 5 \times 10^{-3}/\text{kb}$
- *Neisseria gonorrhoeae* and *N. meningitidis*: 1891/2.18 Mb $\sim 0.9/\text{kb}$
- *Pasteurella multocida*: 927/2.26 Mb $\sim 0.4/\text{kb}$
- *Actinobacillus actinomycetemcomitans*: 1760/4.50 Mb $\sim 0.4/\text{kb}$

Some USS issues

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


Uptake

Do not uptake

 - 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 in gene slightly reduces conservation of embedding site

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

An Agent-Based Model for emergence of USS

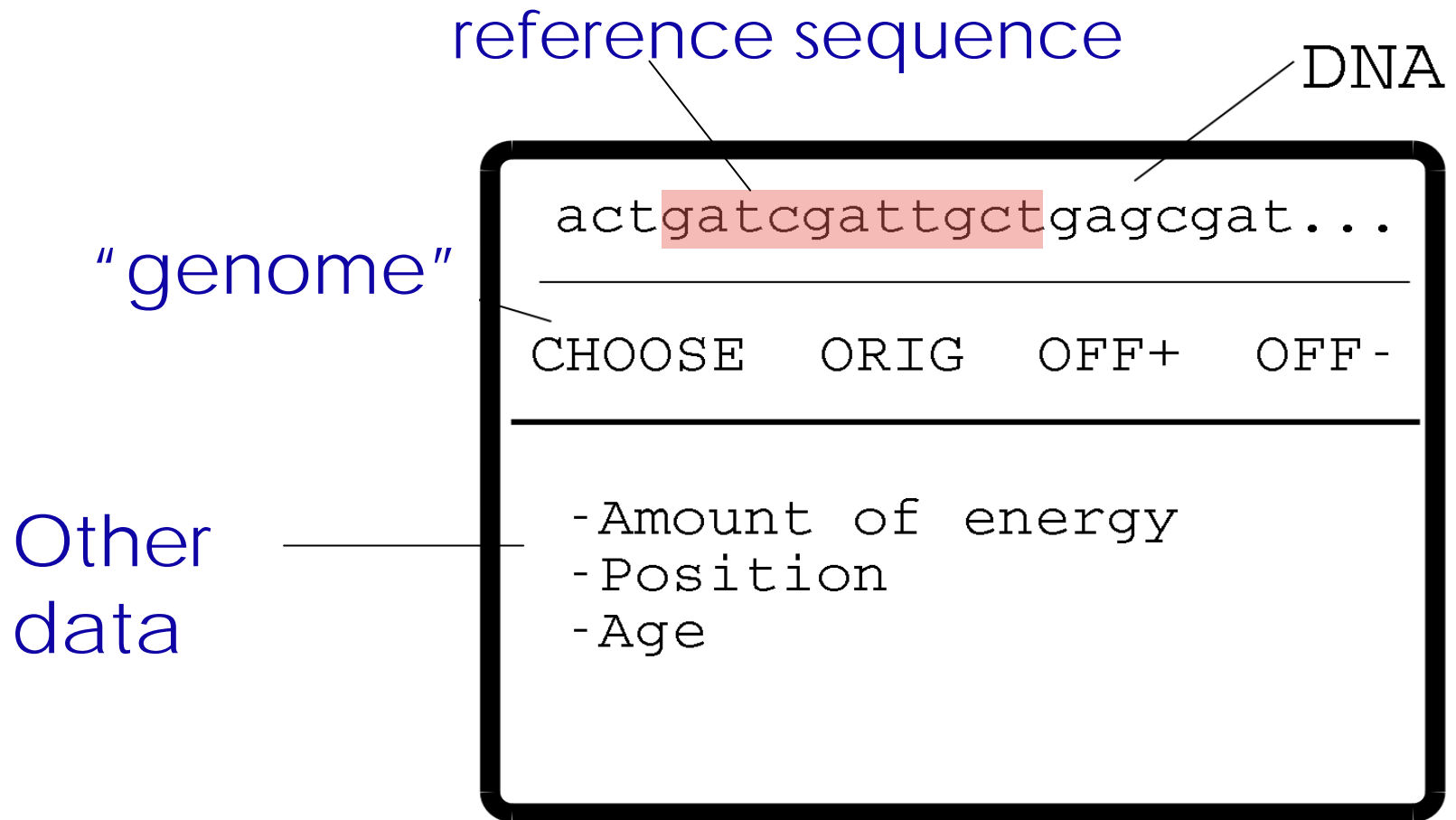
- Uptake of conspecific DNA beneficial
- Uptake of alien DNA not detrimental
- Alien DNA is random
- Initial conspecific DNA is random as well
- Agents must learn to distinguish between conspecific and alien DNA

Structure of the Model

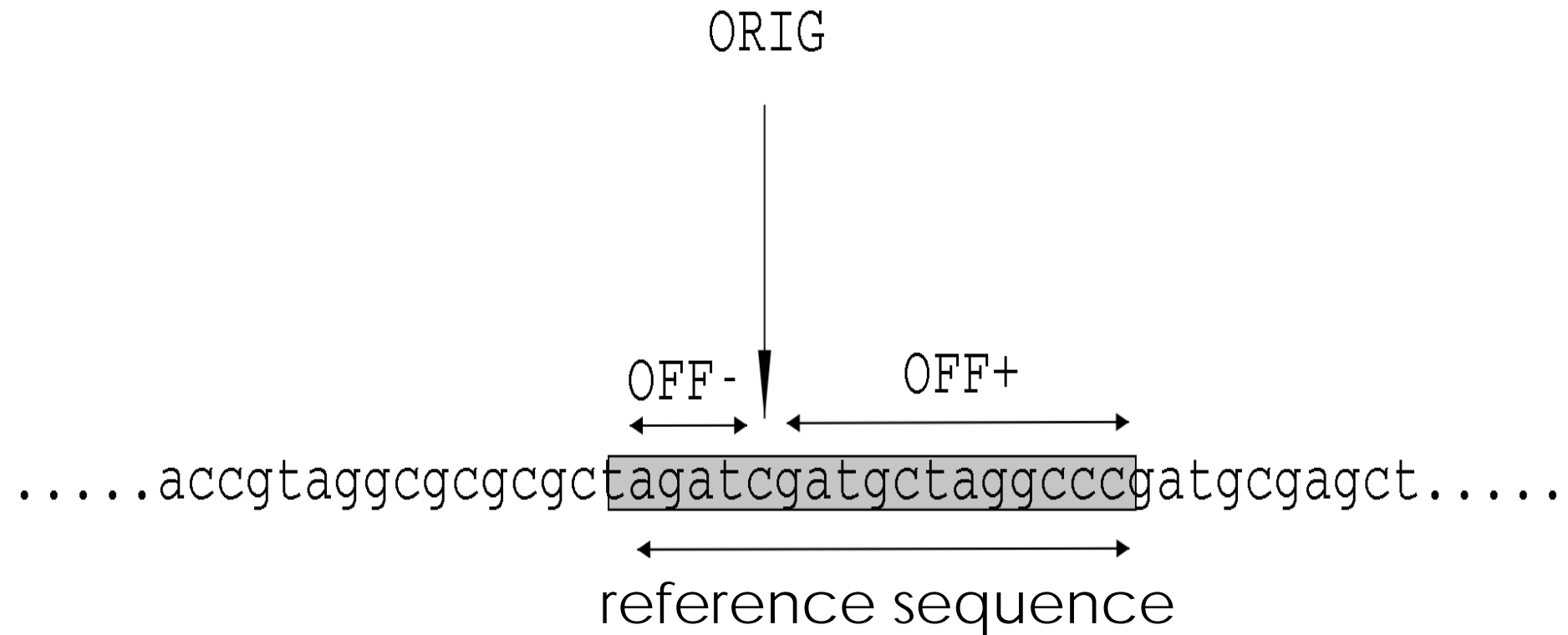
- Agents (the genomes)
- Environment
- Rules

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

a site



Agents (>1 per site)
bacterial/alien DNAs
“..cggtgactgaac...”

Two types of DNA fragments

..aacggtgcctatcgt..
..ttcacgtggtgactc..

} "alien"

..atccgcgcggtttacg..
..aattttacacaggcg..

} "bacterial"

Reference
sequence

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
 - $\frac{1}{2}$ 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

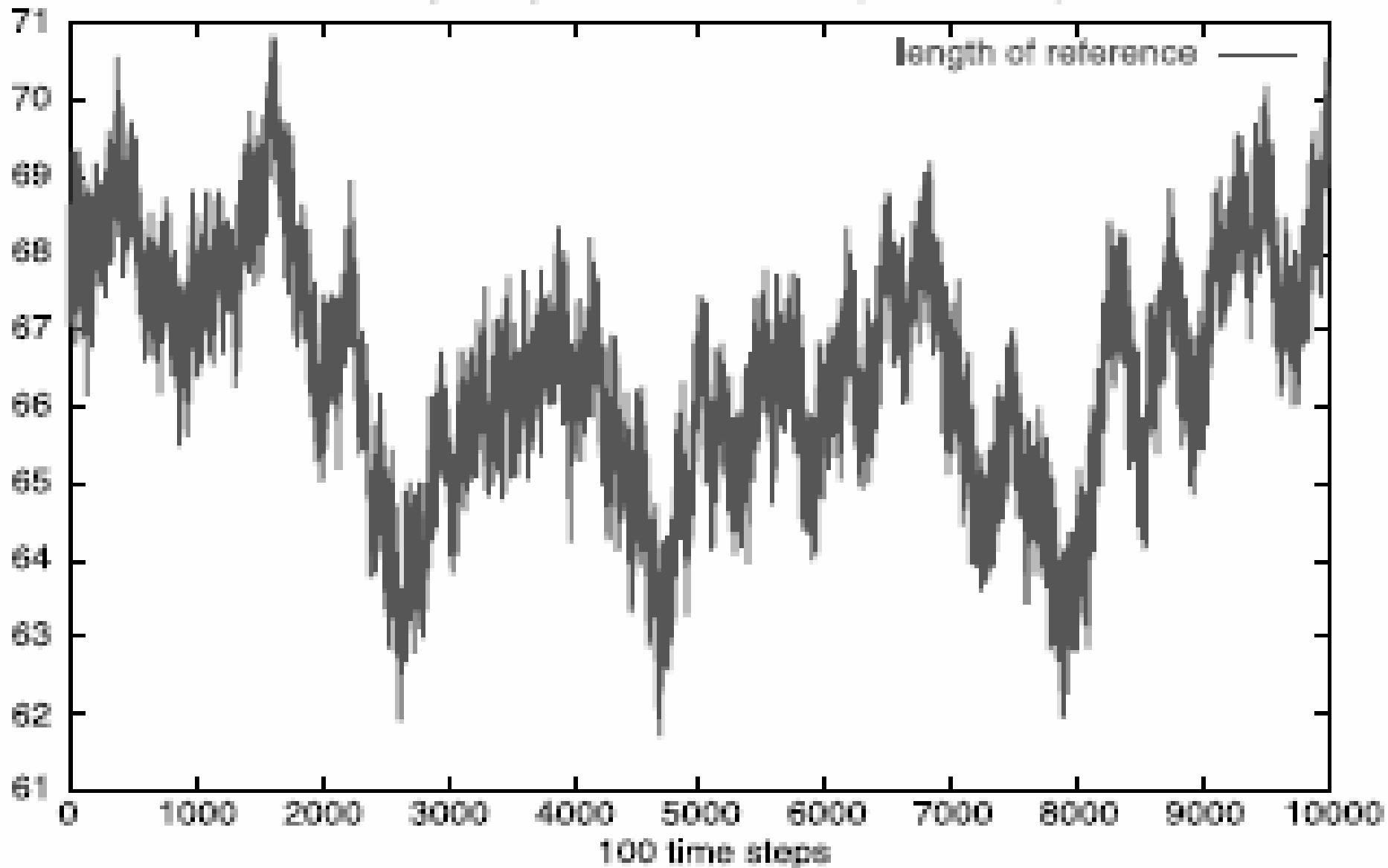
Number of updates = 1,000,000

Energy from alien fragment = 1

Energy from bacterial fragment = 2

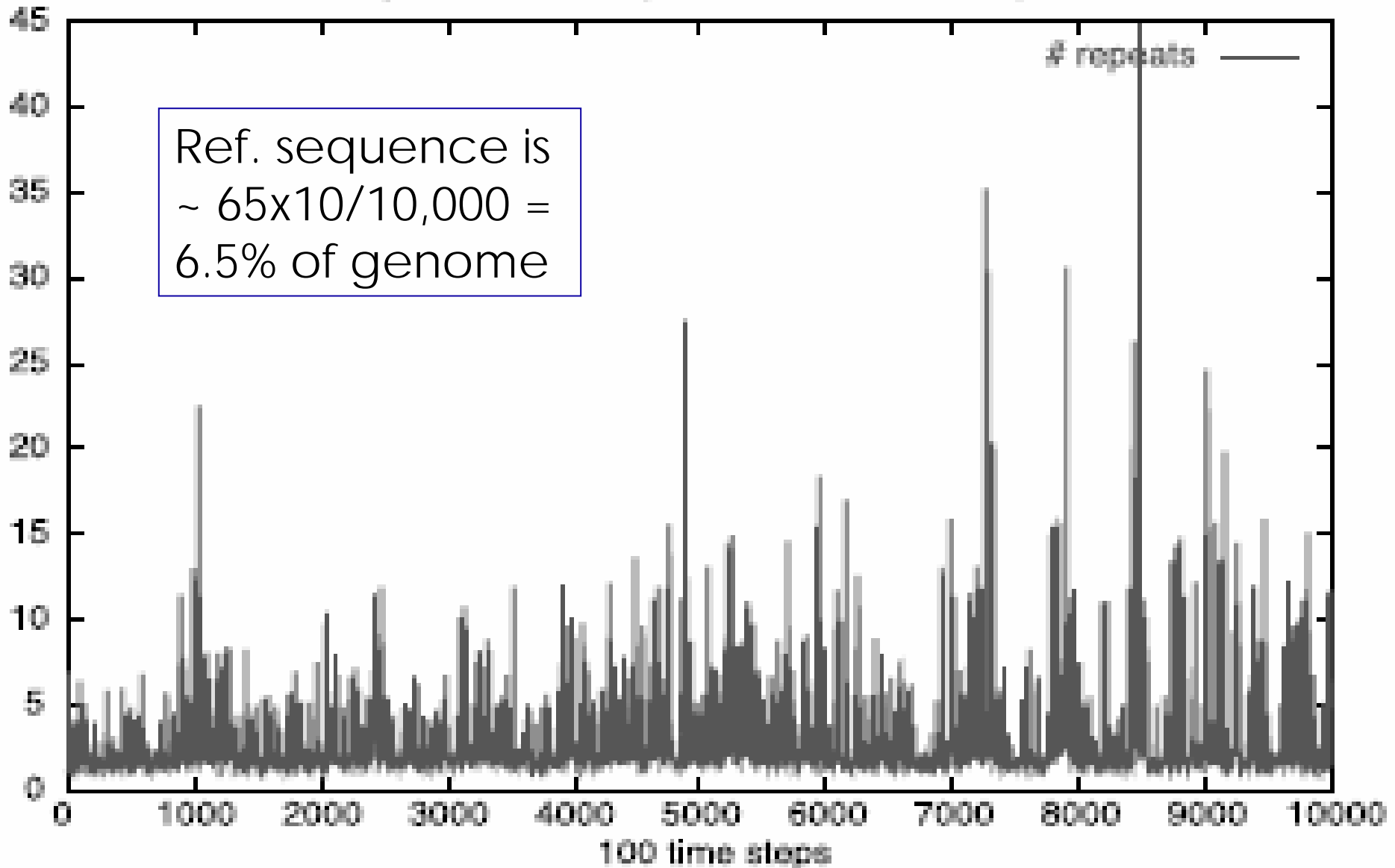
Nothing happens

Average length of reference sequences

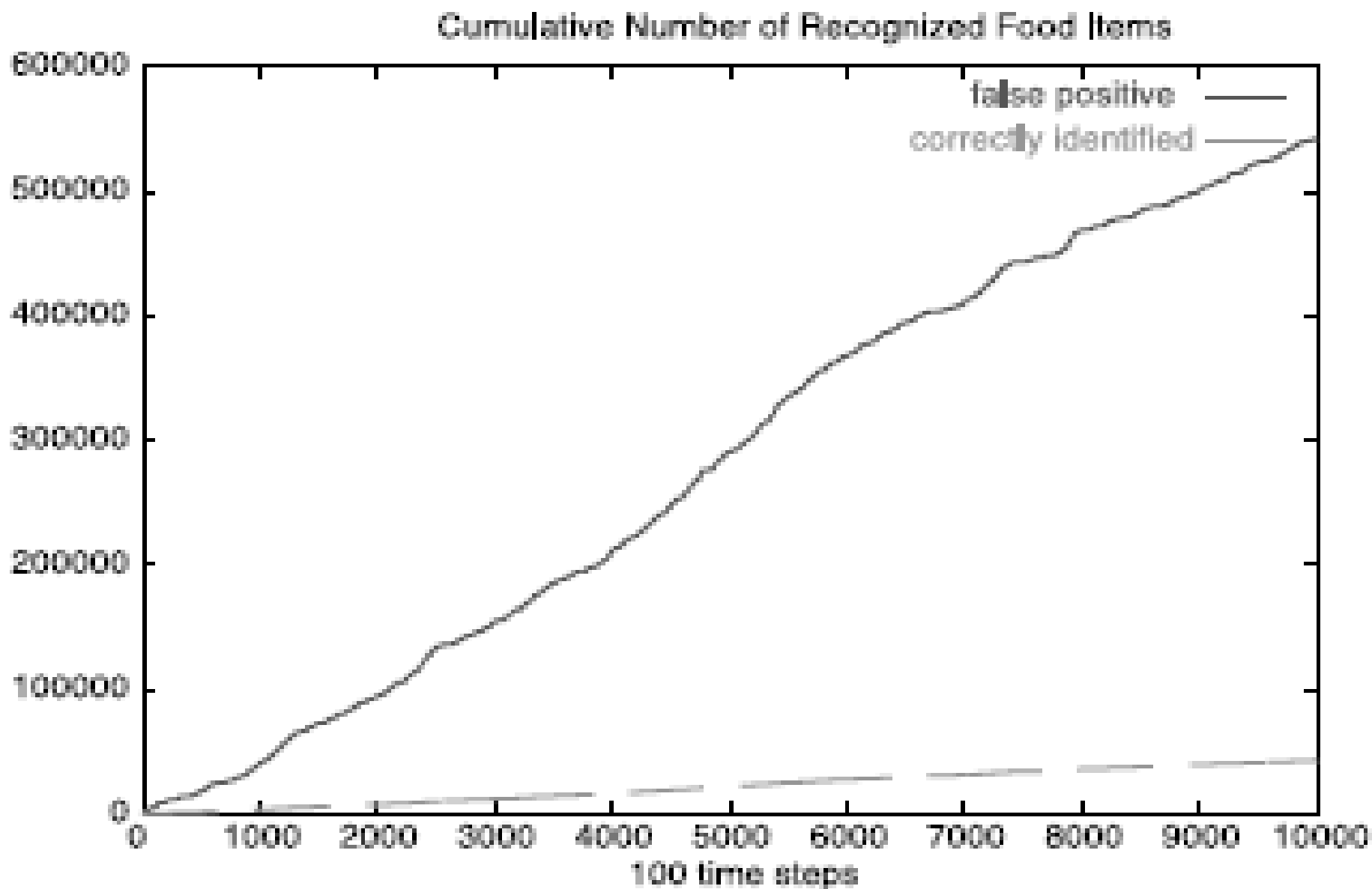


Average number of repeats of reference sequences on DNA

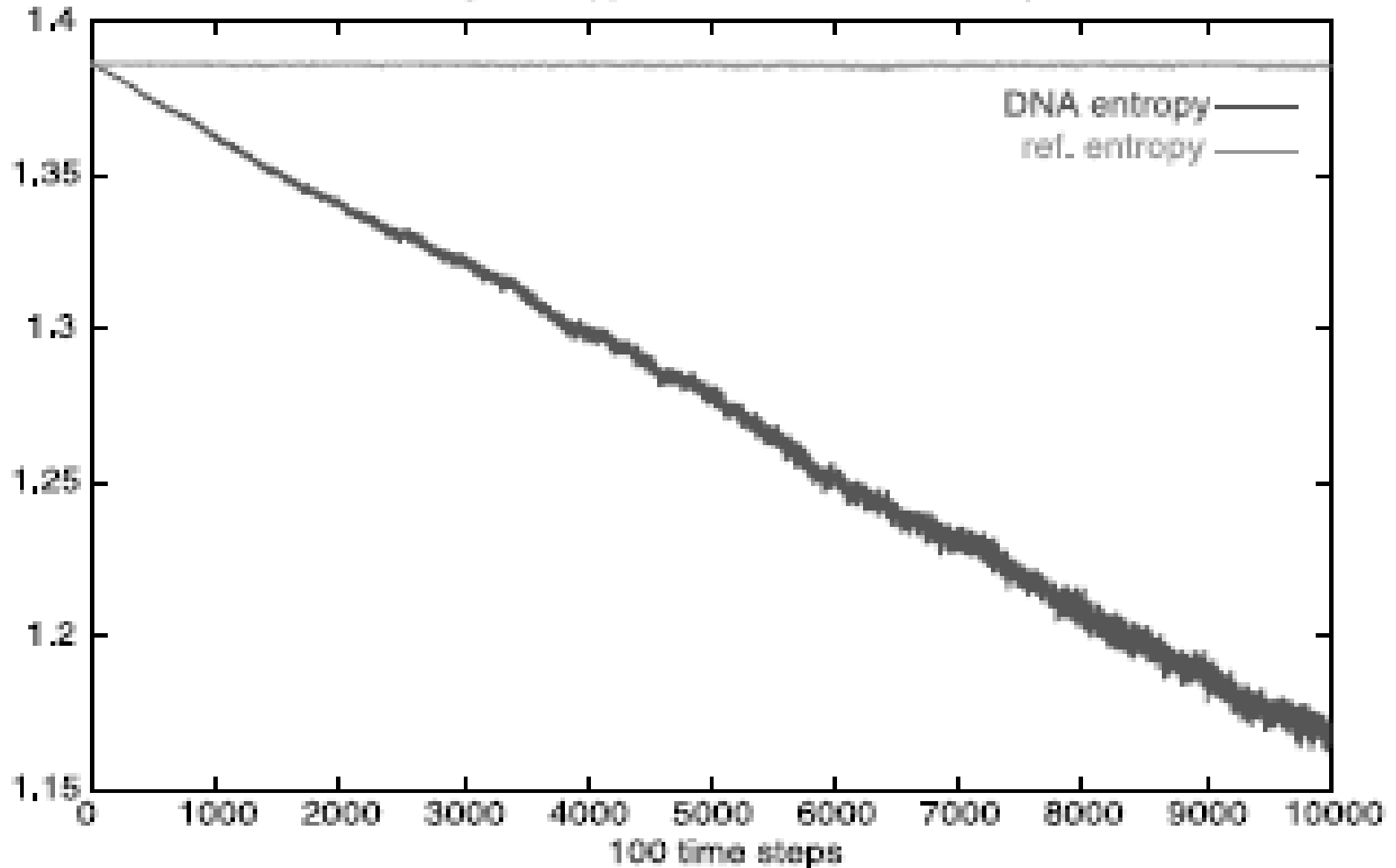
Ref. sequence is
 $\sim 65 \times 10 / 10,000 =$
6.5% of genome



Cumulative number of recognized uptakes



Average entropy of DNA/reference in population



Run 2

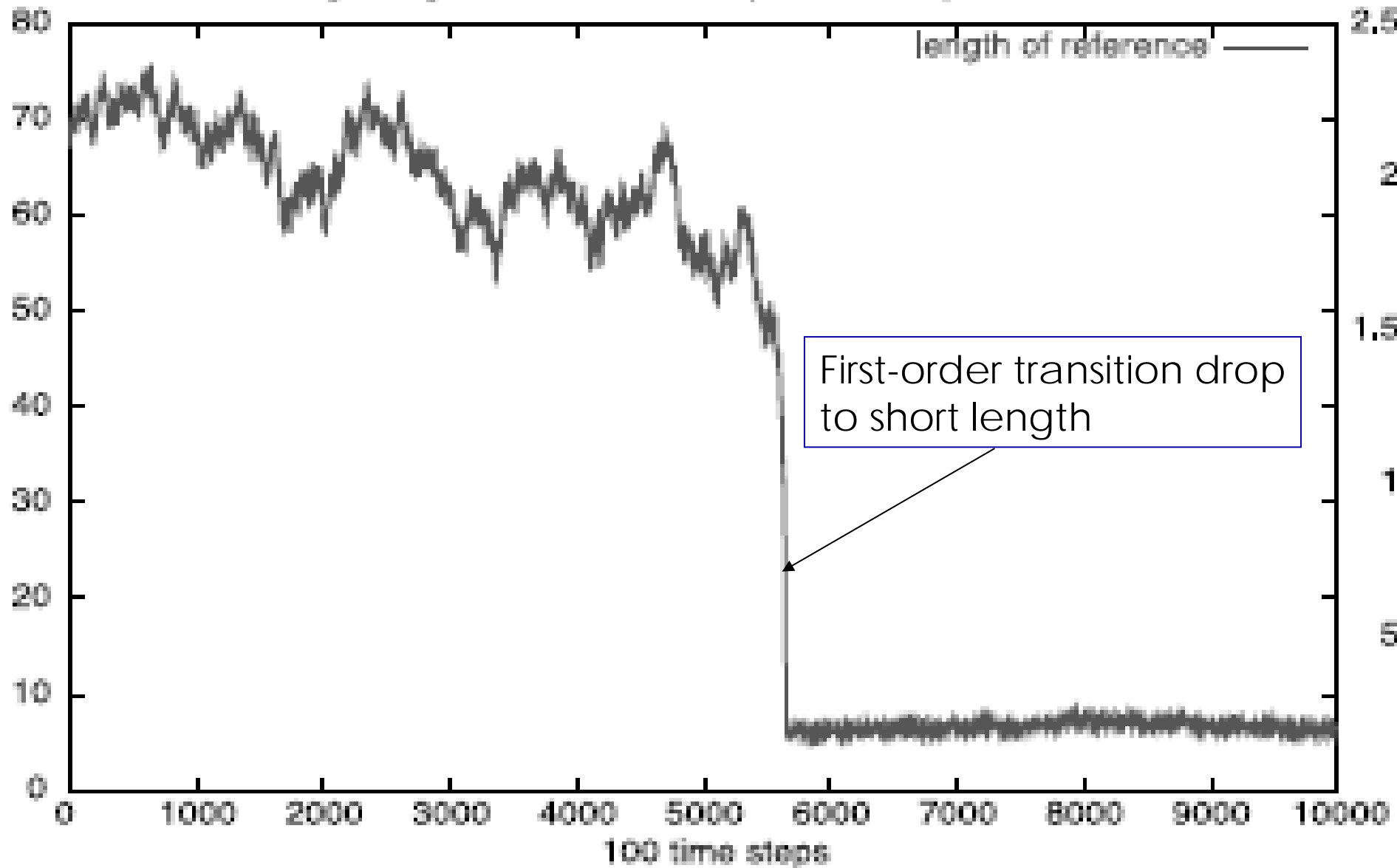
Number of updates = 1,000,000

Energy from alien fragment = 1

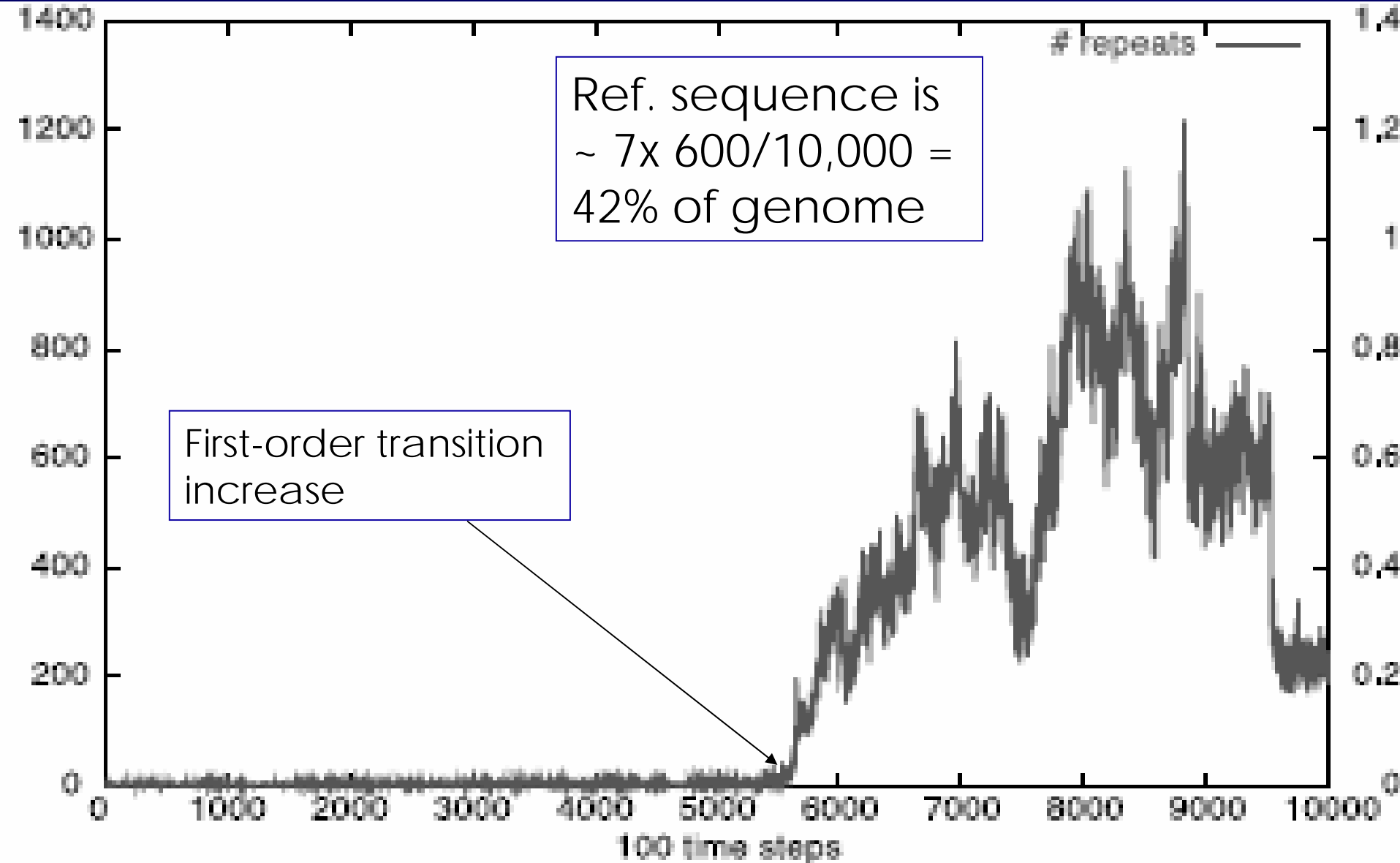
Energy from bacterial fragment = 3

Spontaneous emergence of USS

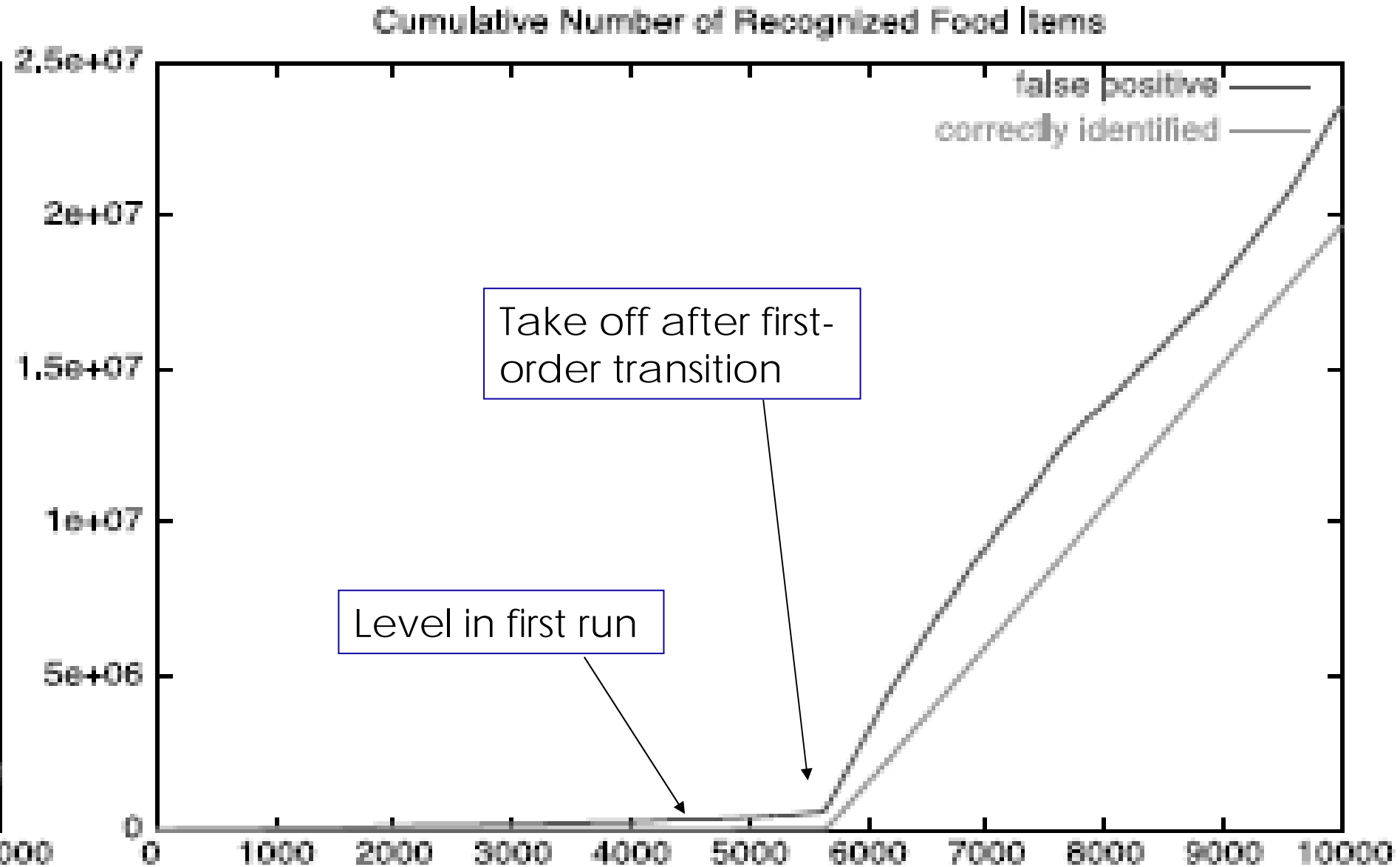
Average length of reference sequences



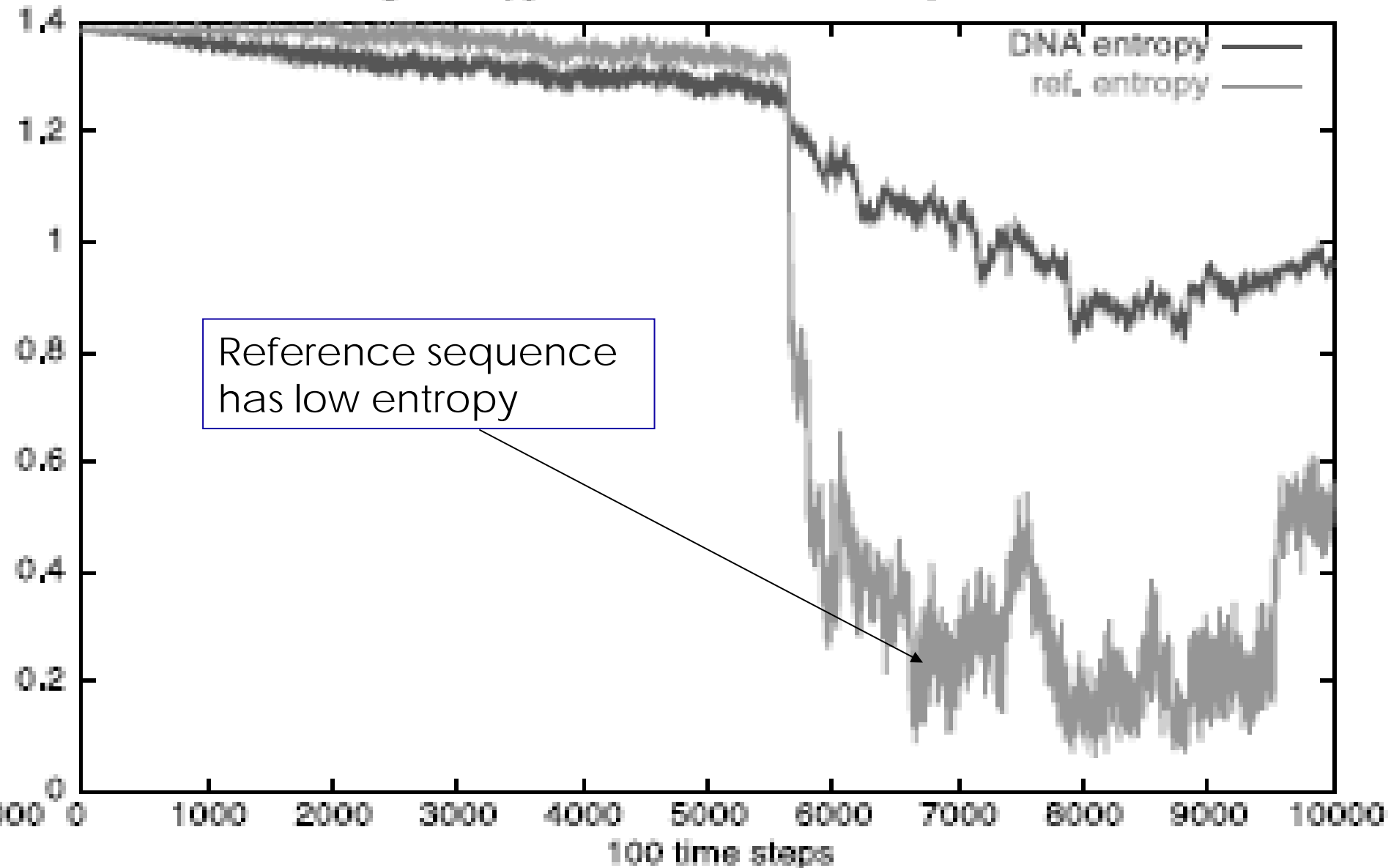
Average number of repeats of reference sequences on DNA



Cumulative number of recognized uptakes



Average entropy of DNA/reference in population



Summary

- First CA with “realistic” genomes as agents
- Demonstrated emergence of USS as a **first-order transition**
- Uniformity symmetry was **spontaneously broken**
- Result was far more complex than Wolfram’s “complexity” work

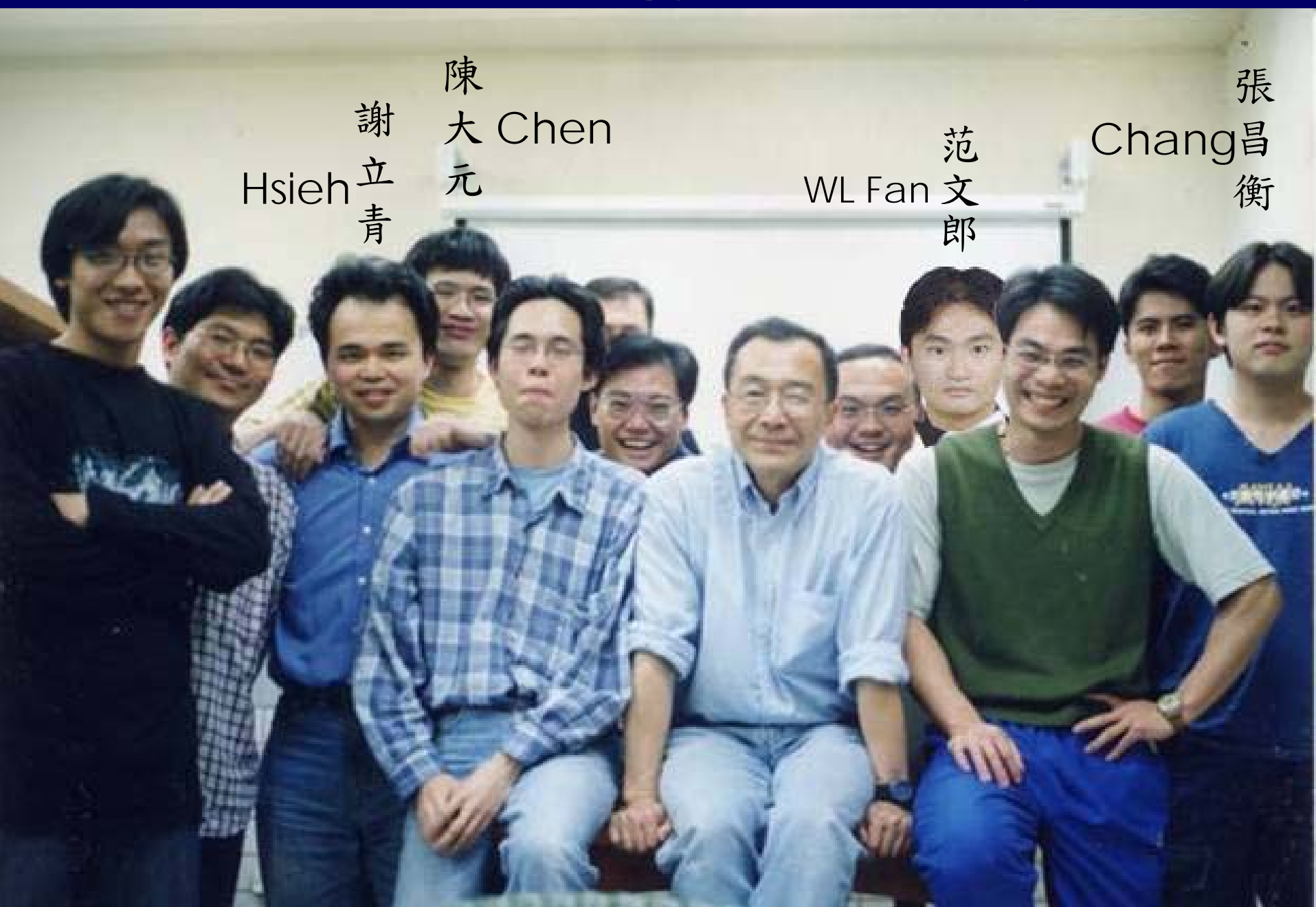
Outlook

- An analytic version of this work
- Prototype of emergence in biology?

People

- Genome growth
 - **Li-Ching Hsieh** 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 - Cellular automata
 - **Dominique Chu/Gross**, Comp. Sci., U. Manchester, UK
 - **Tom Lenaerts**, U. Libre de Bruxelles, Brussels, Belgium

Computation Biology Laboratory (2003)



謝立青
Hsieh

陳大元
Chen

范文郎
WL Fan

張昌衡
Chang

Our papers are found at
Google: HC Lee

Thank you!