Emergence and Spontaneous Symmetry Breaking in Genome Evolution by *Cellular Automata*

Workshop on
Critical Phenomena & Complex Systems
Chinese Cultural University
2006 March 31-April 1

HC Lee
Department of Physics
Grad. Inst. Systems Biology & Bioinformatics
National Central University

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

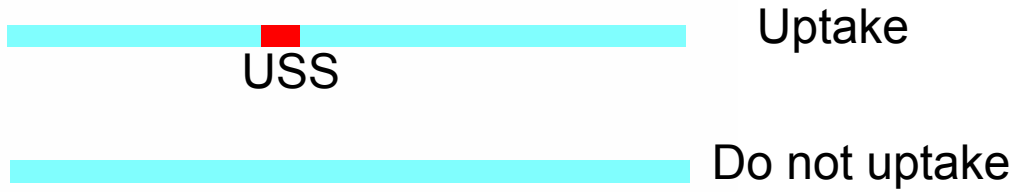
- The DNA of some naturally "competent" species of bacteria contains a large number of evenly distributed copies of a perfectly conserved short sequence.
- This highly overrepresented sequence is believed to be an uptake signal sequence (USS) that helps bacteria to take up DNA selectively from (dead) 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

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)

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

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.

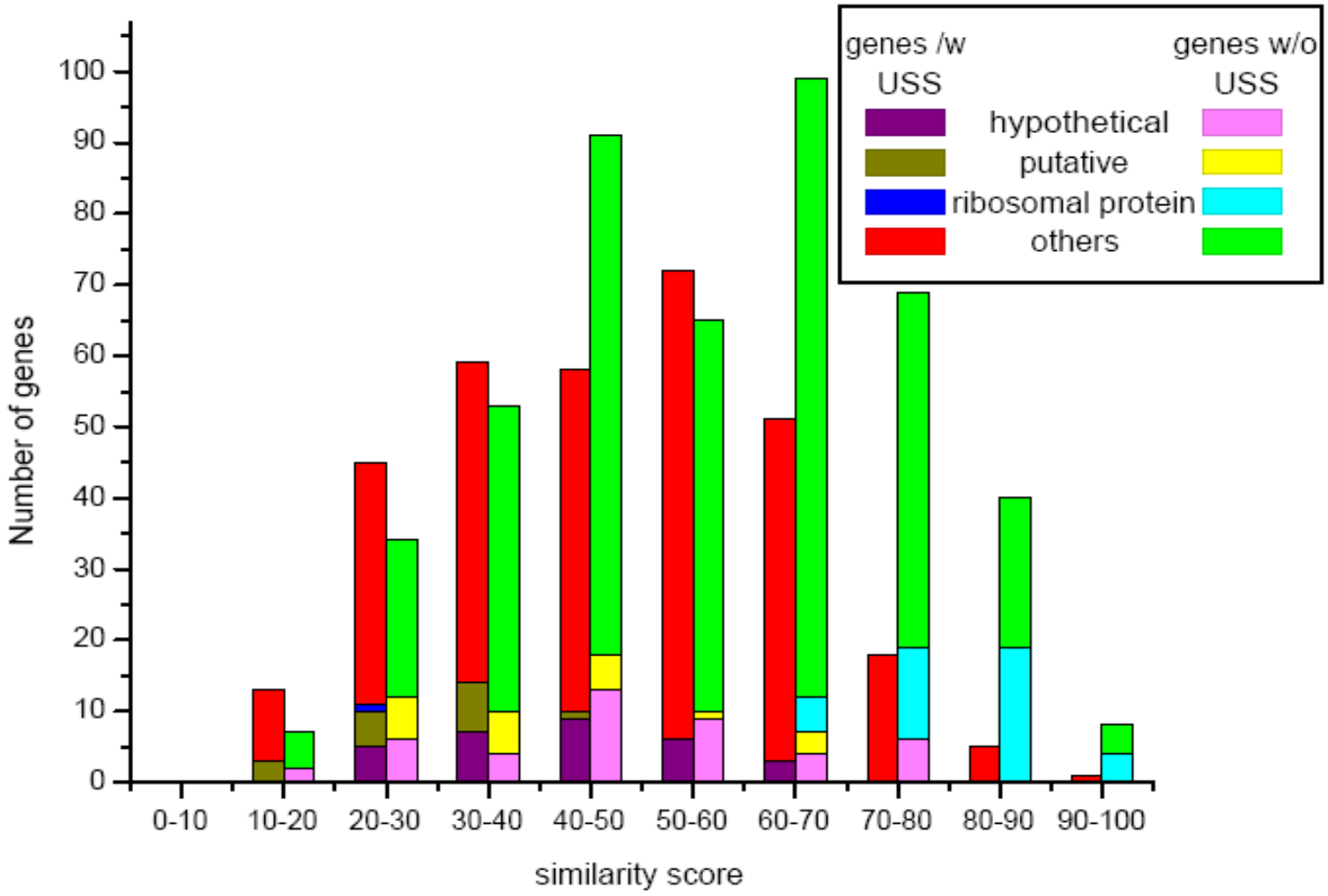
2 Including cases where both genes are in the same strand and cases where one gene in plus strand and the other in minus strand.

* 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

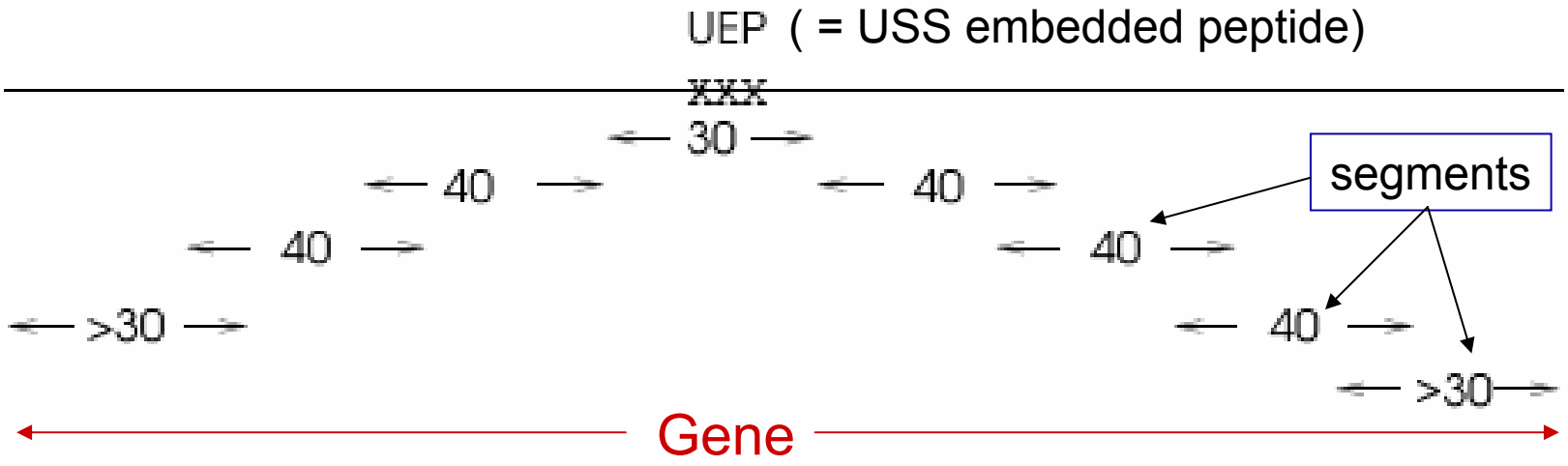
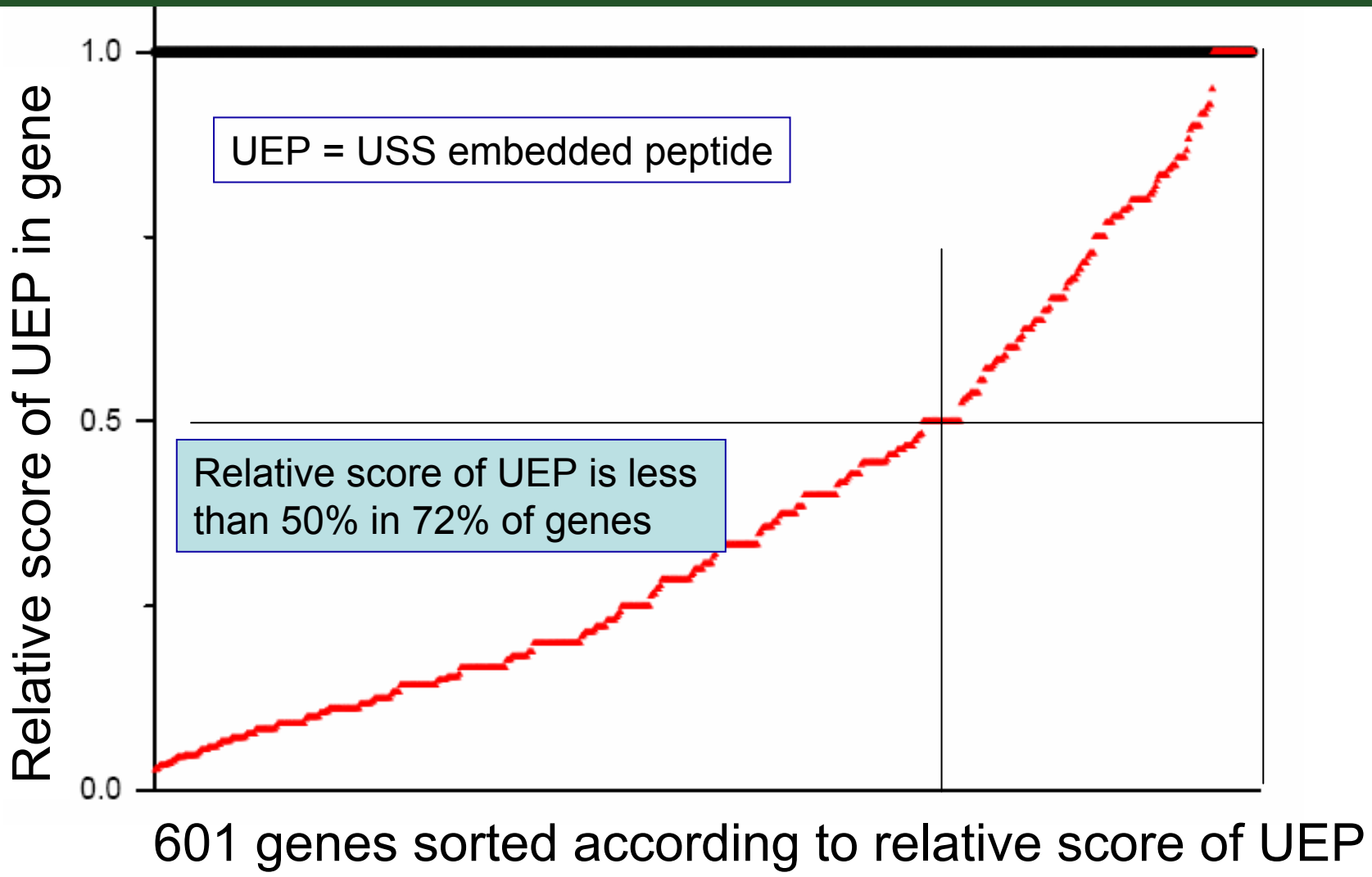


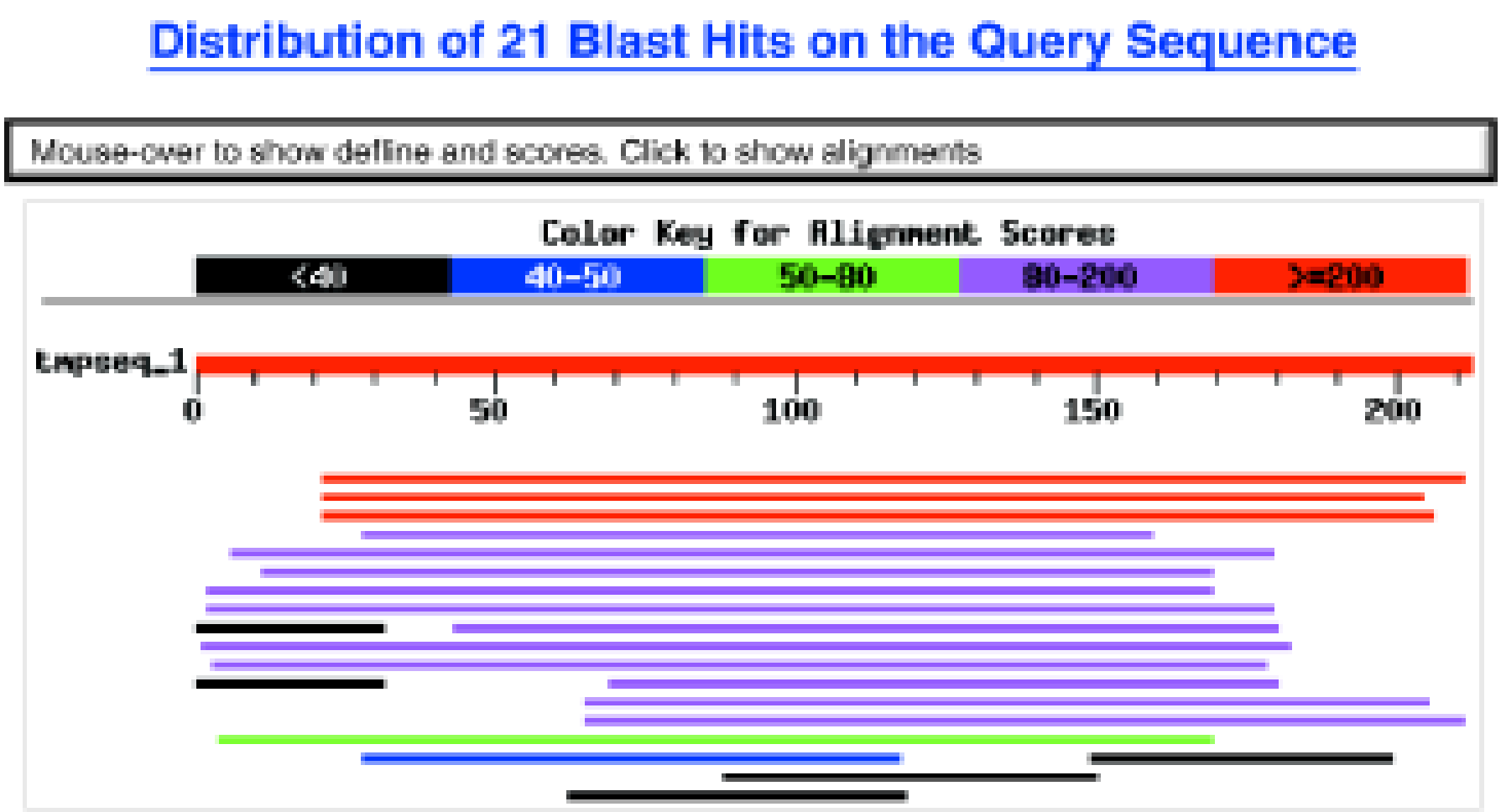
Figure 1: Segmentation of a protein sequence. XXX is the position of the UEP in the query, or of the corresponding peptides in a match.

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



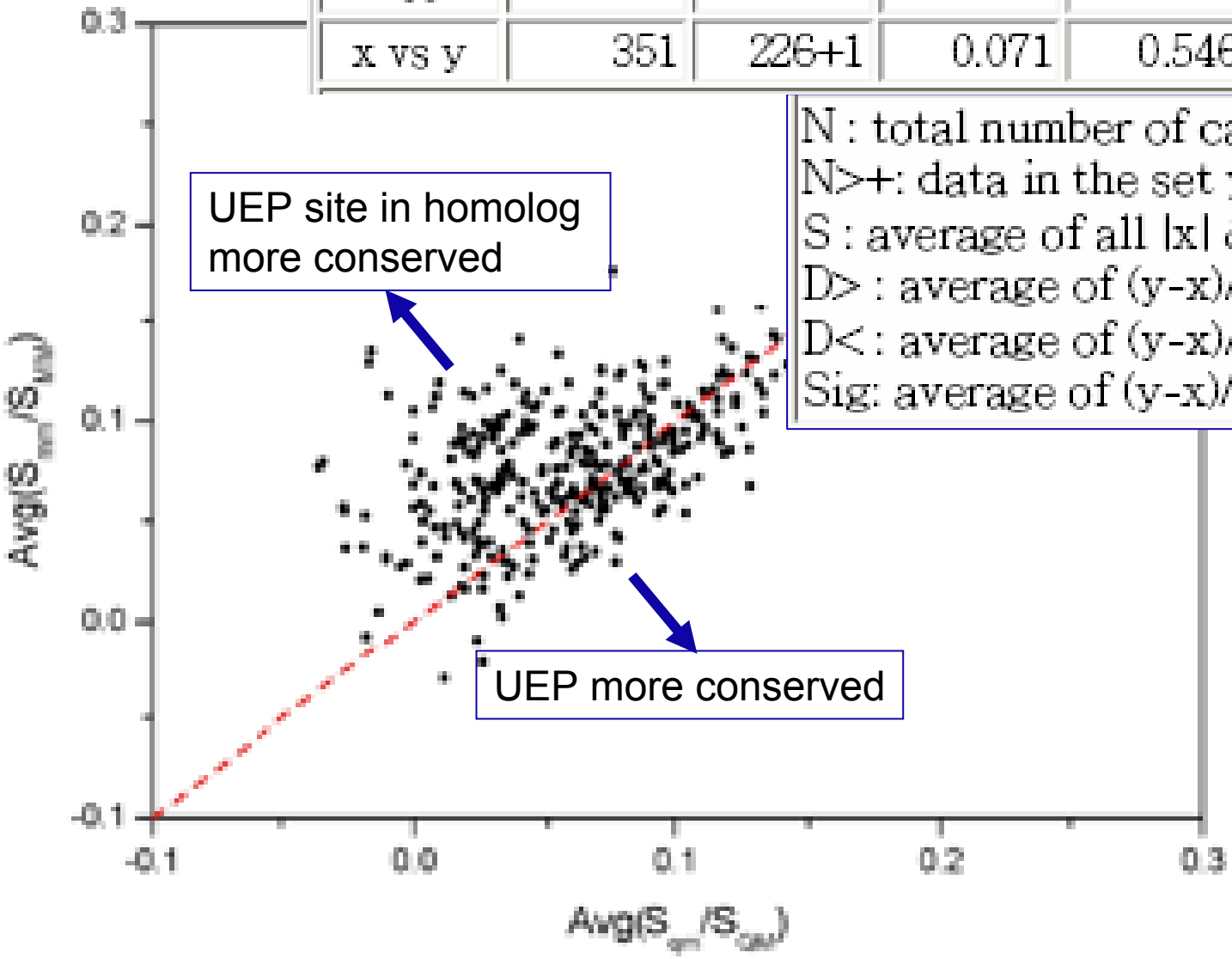
Homolog search of UEP embedded gene by BLAST



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

Our central Assumption

Uptake of conspecific DNA is more beneficial than uptake of other DNA.

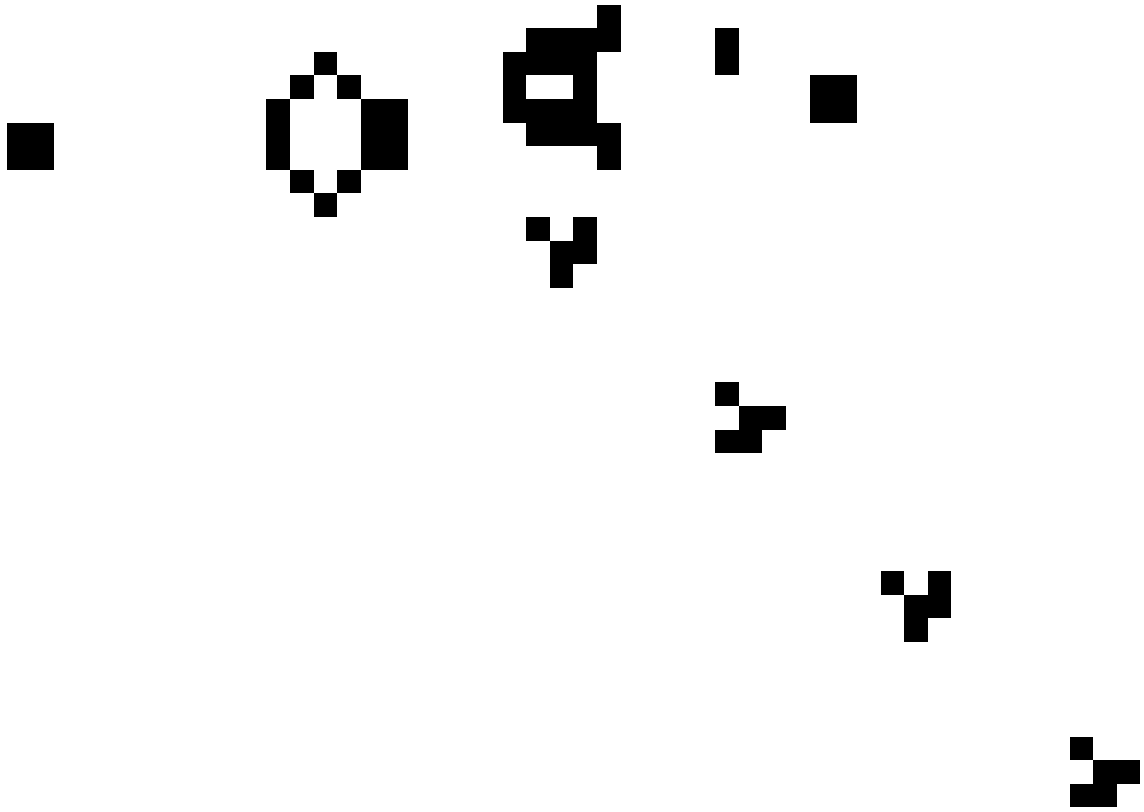
Can we demonstrate the emergence of USS in a computational model?

Agent-Based Model (Cellular automata)

- Reality is complex, but models don't have to 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

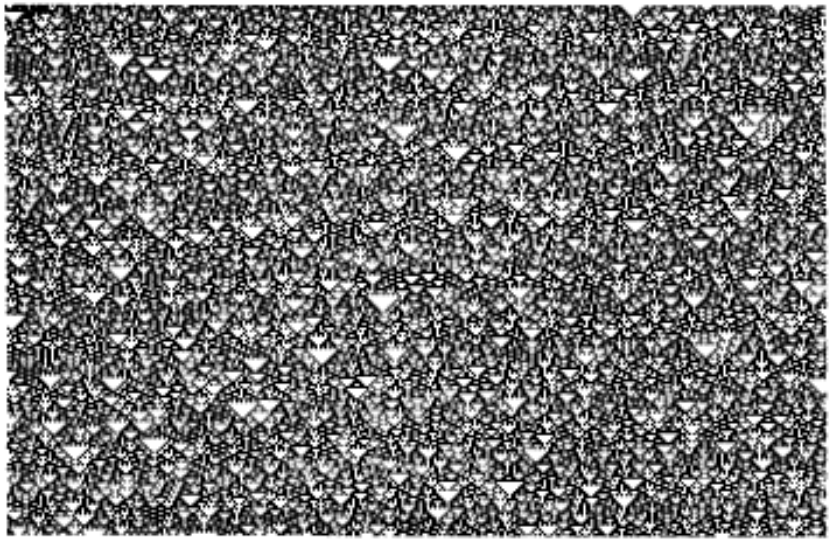
Conway's Game of Life: Simple rules may generate not-so-simple patterns

Gospers glider gun

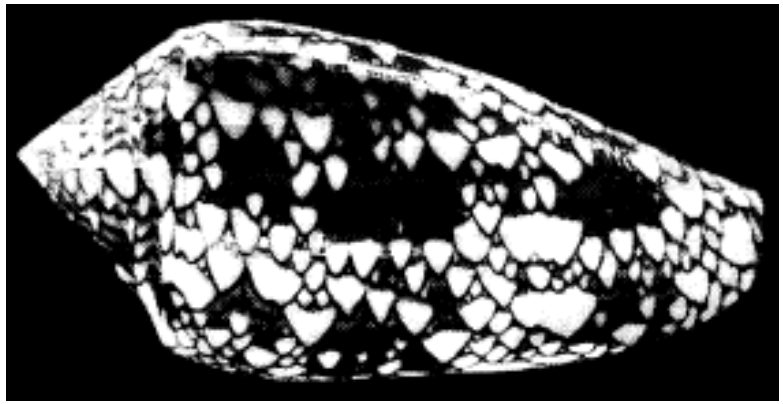


Wolfram: Self-Organization in Cellular Automata

Cellular Automata



Mollusk shell



An Agent-Based Model for emergence of USS

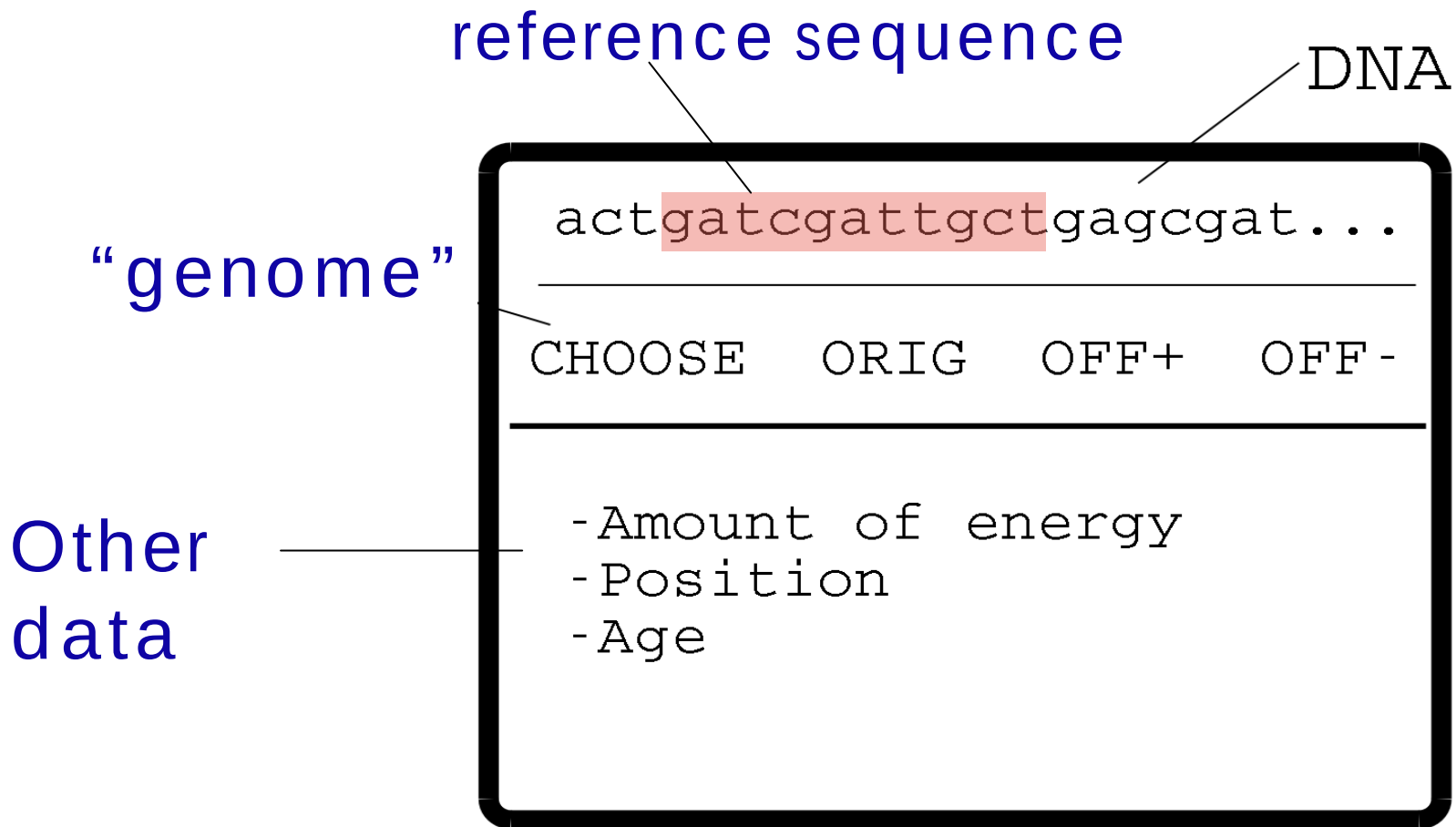
- Uptake of conspecific DNA beneficial
- Uptake of alien DNA not detrimental
- Alien DNA is random
- Initial conspecific DNA is random as well
  - i.e. reference sequence is random
- Agents must learn to distinguish between conspecific and alien DNA

Structure of the Model

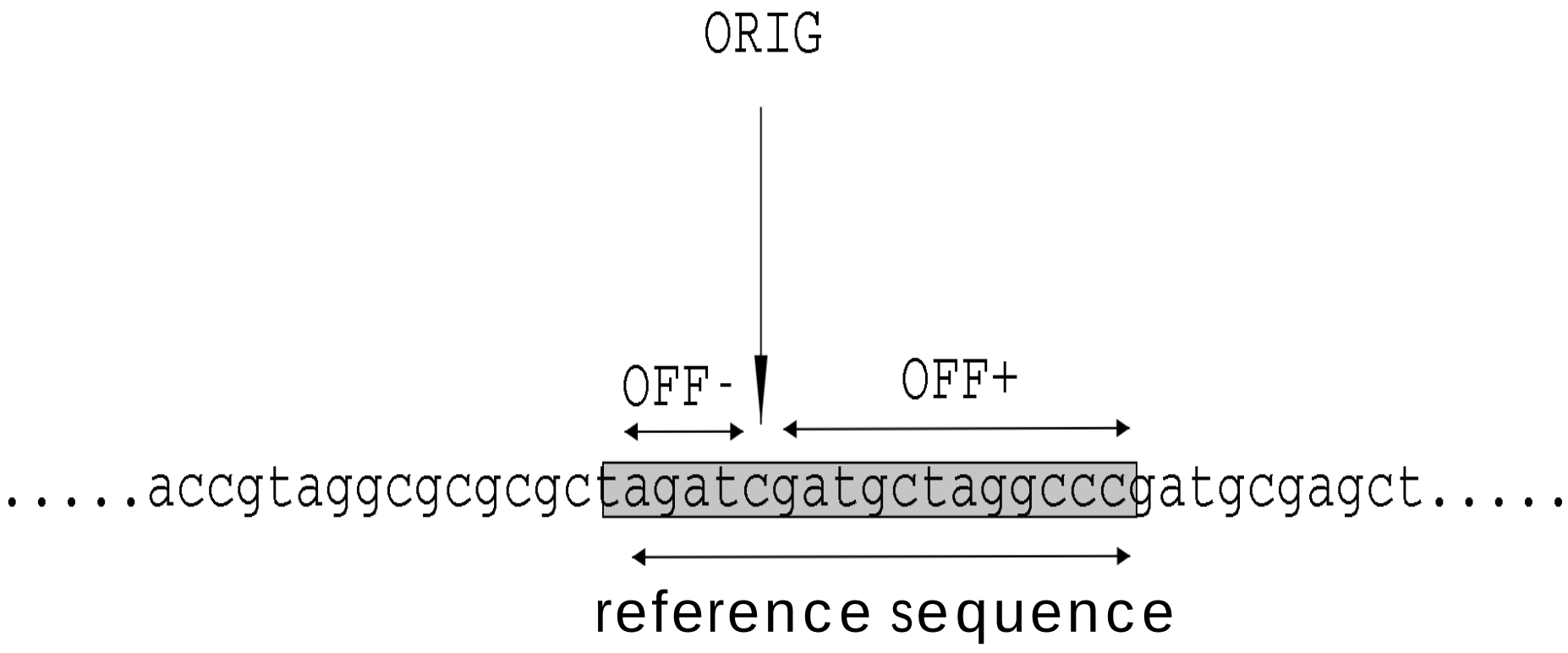
- Agents (the genomes)
  - Compare: agents in Conway's and Wolfram's games are just regular geometric shapes (squares or triangles)
- Environment
- Rules

AGENTS

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



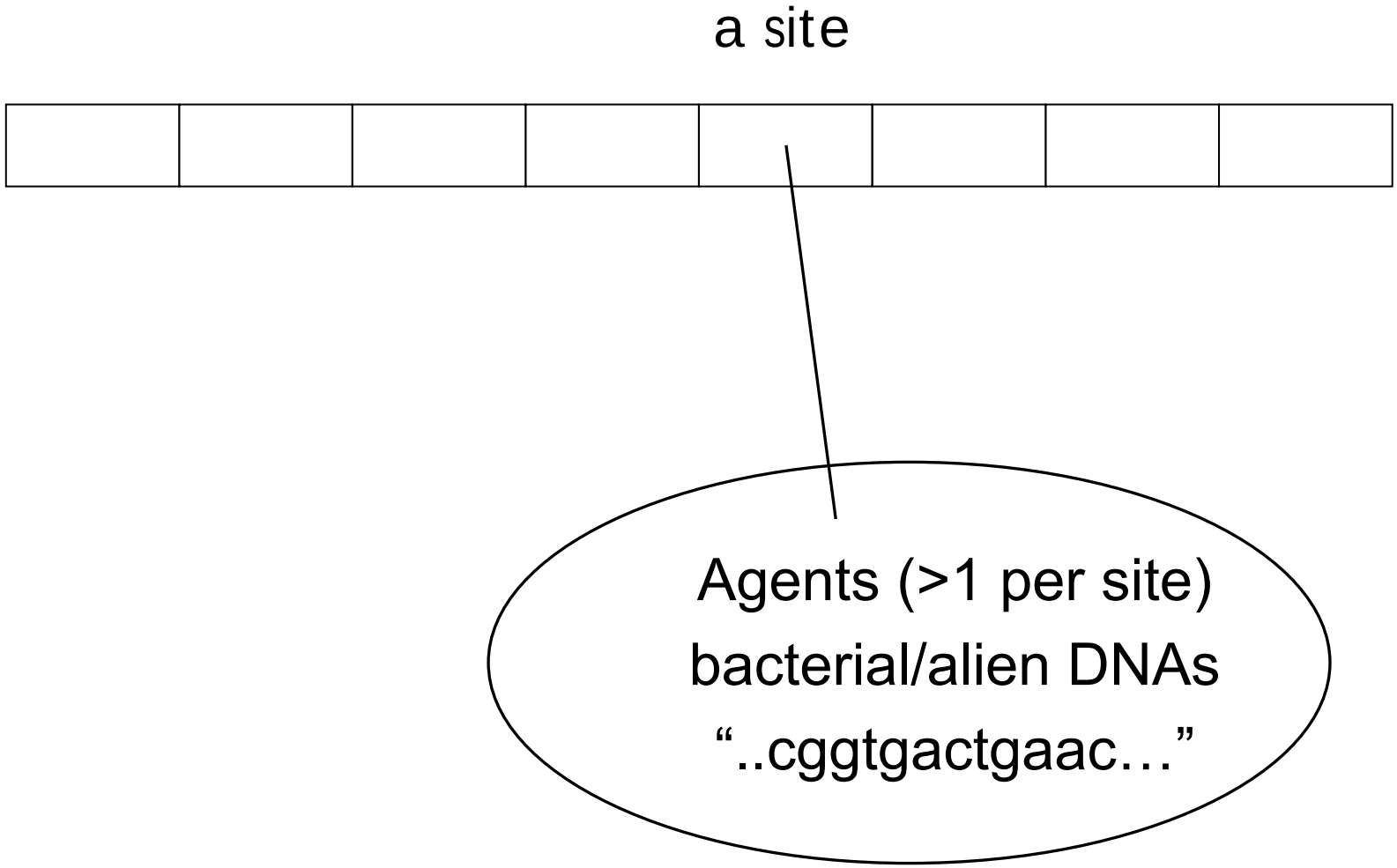
“Genome” contains information pointing to the reference sequence



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").

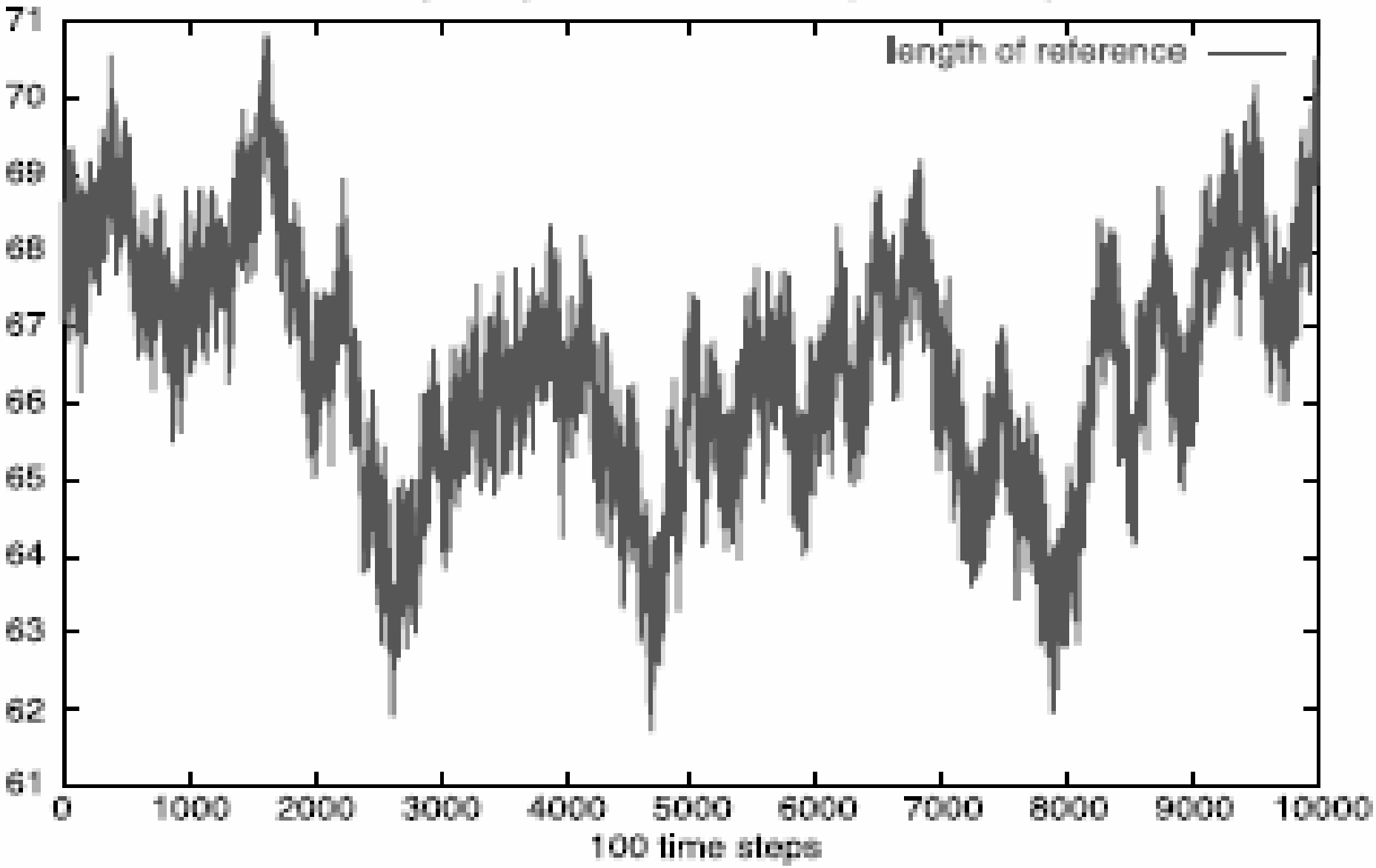
Run parameters

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

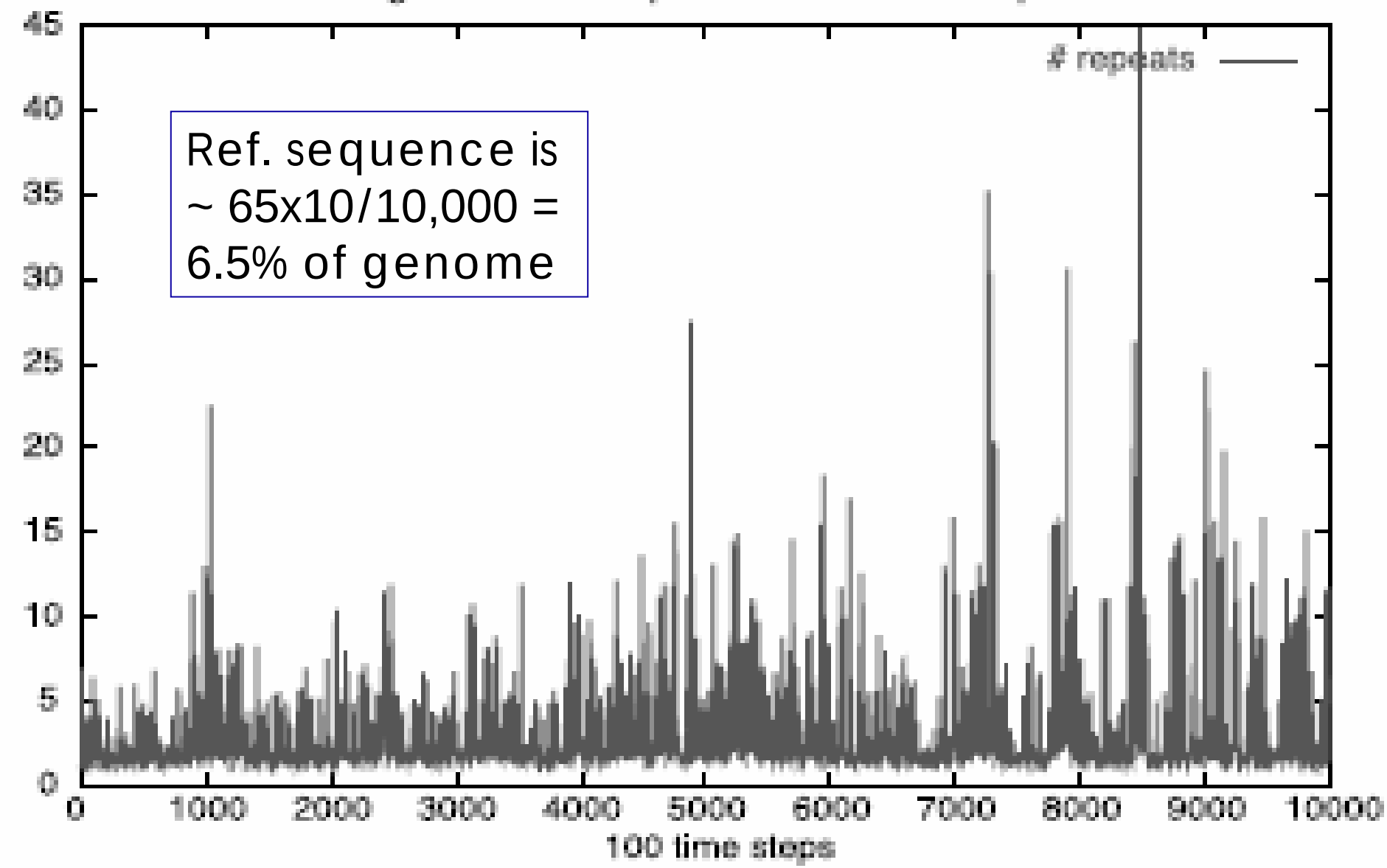
Run 1

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

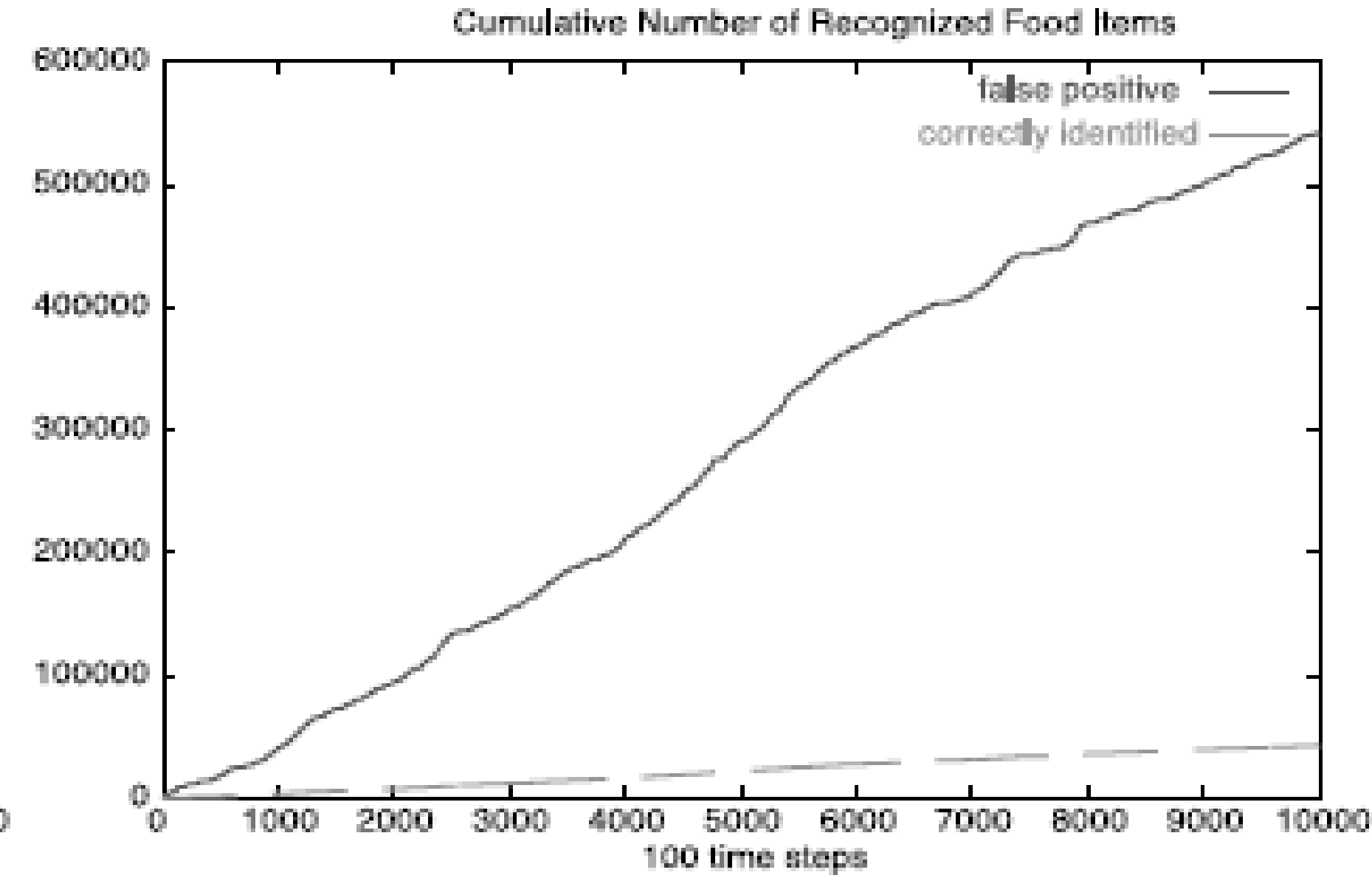
Average length of reference sequences



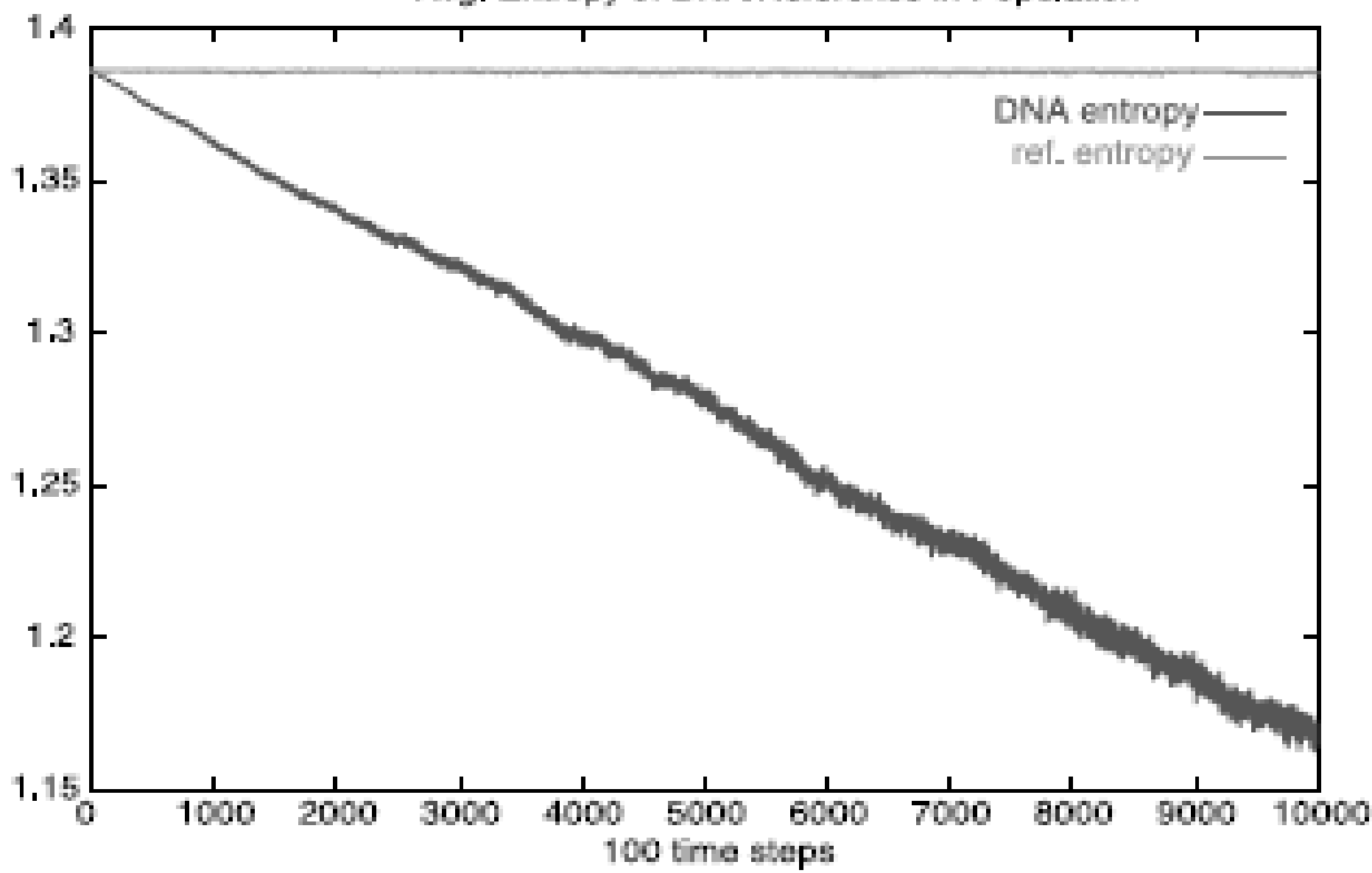
Average number of repeats of reference sequences on DNA



Cumulative number of recognized uptakes



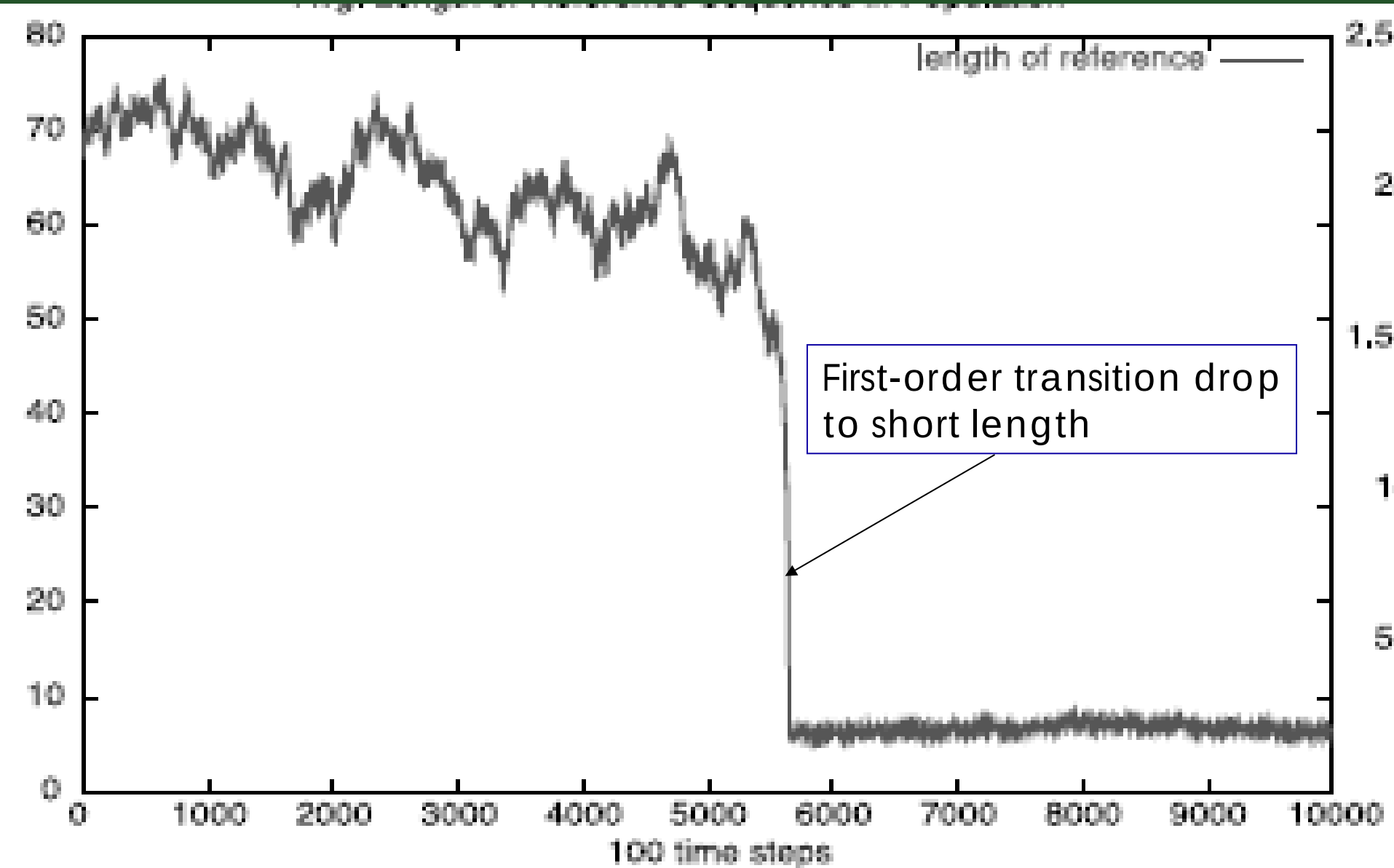
Average entropy of DNA/reference in population



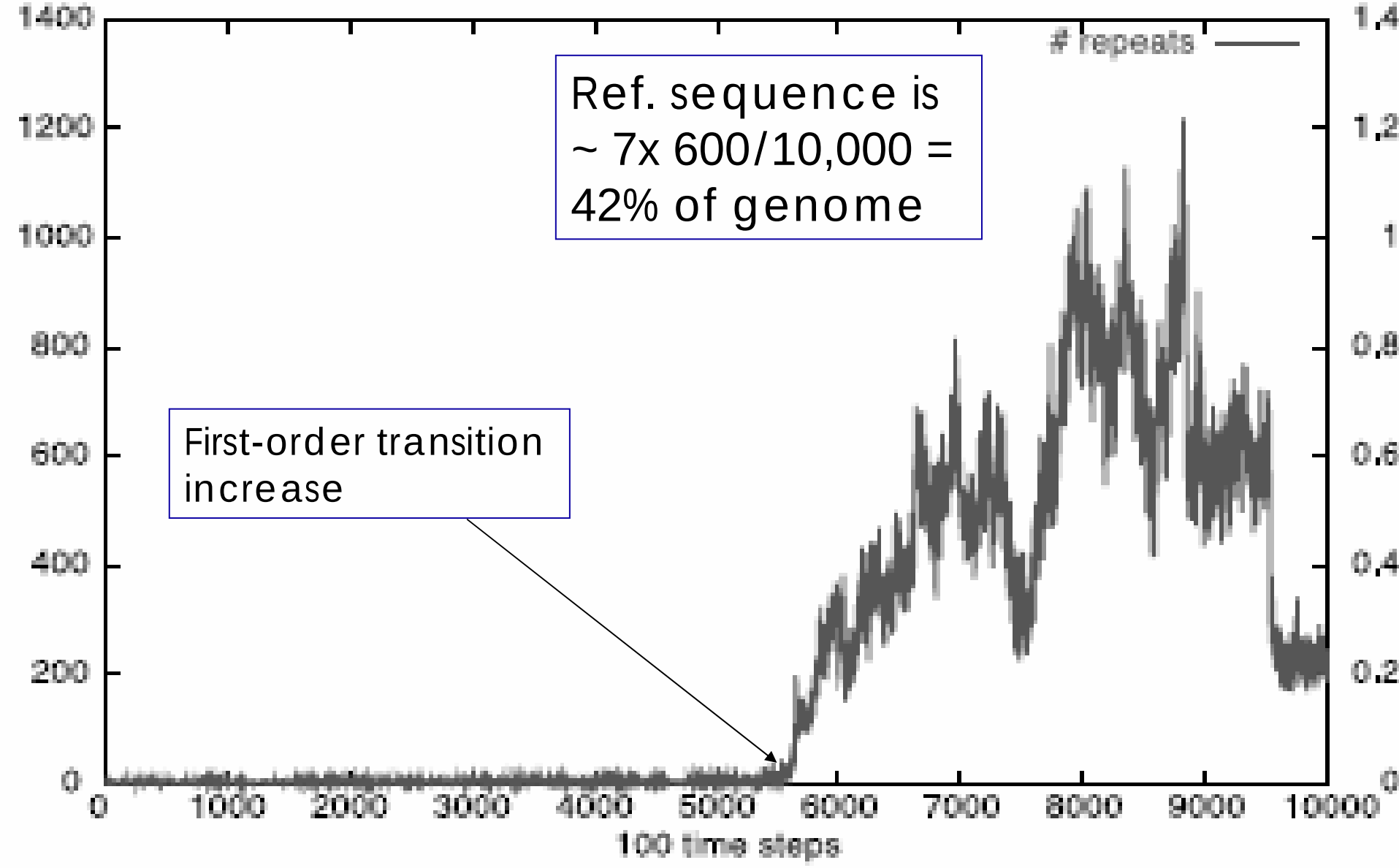
Run 2

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

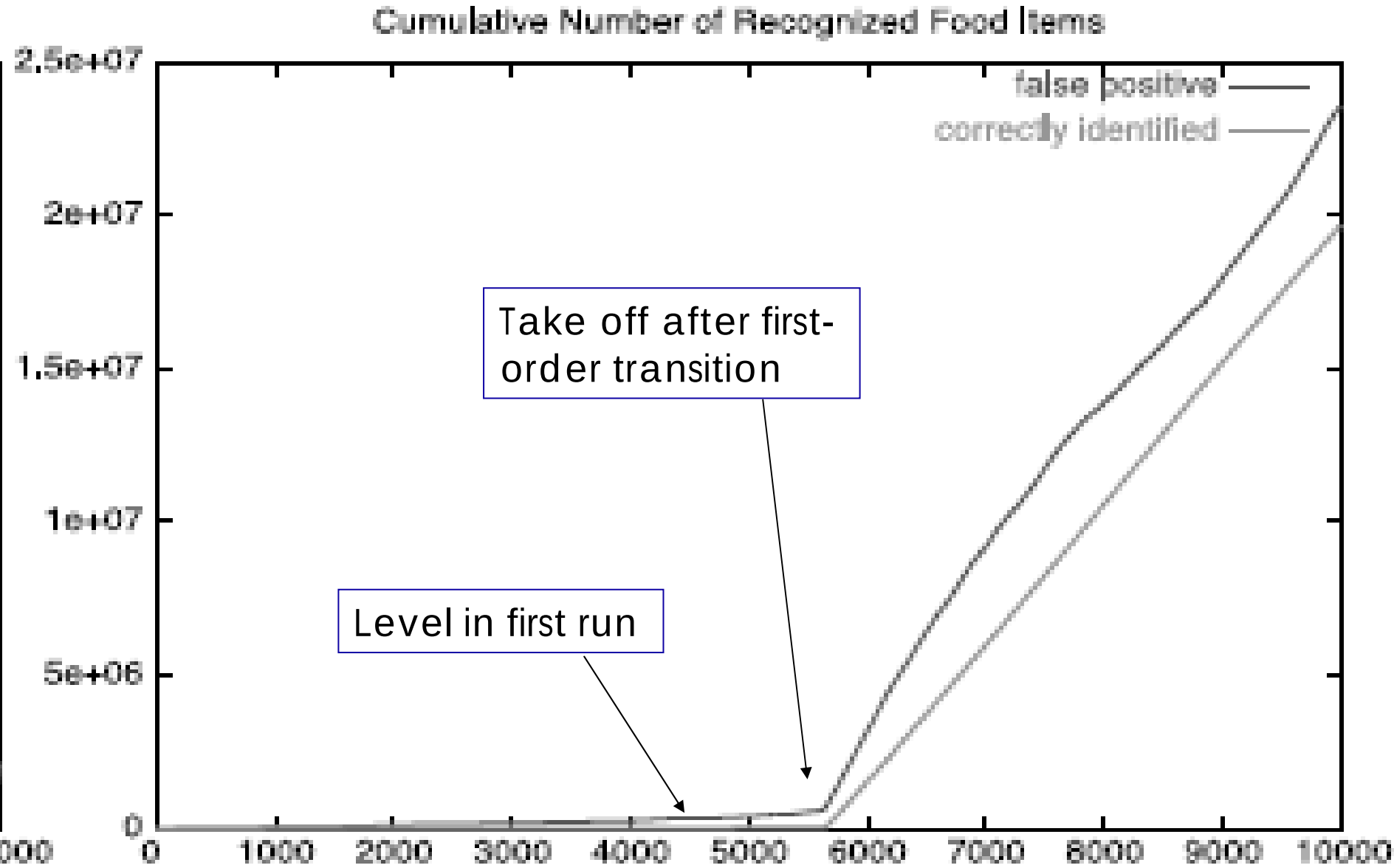
Average length of reference sequences



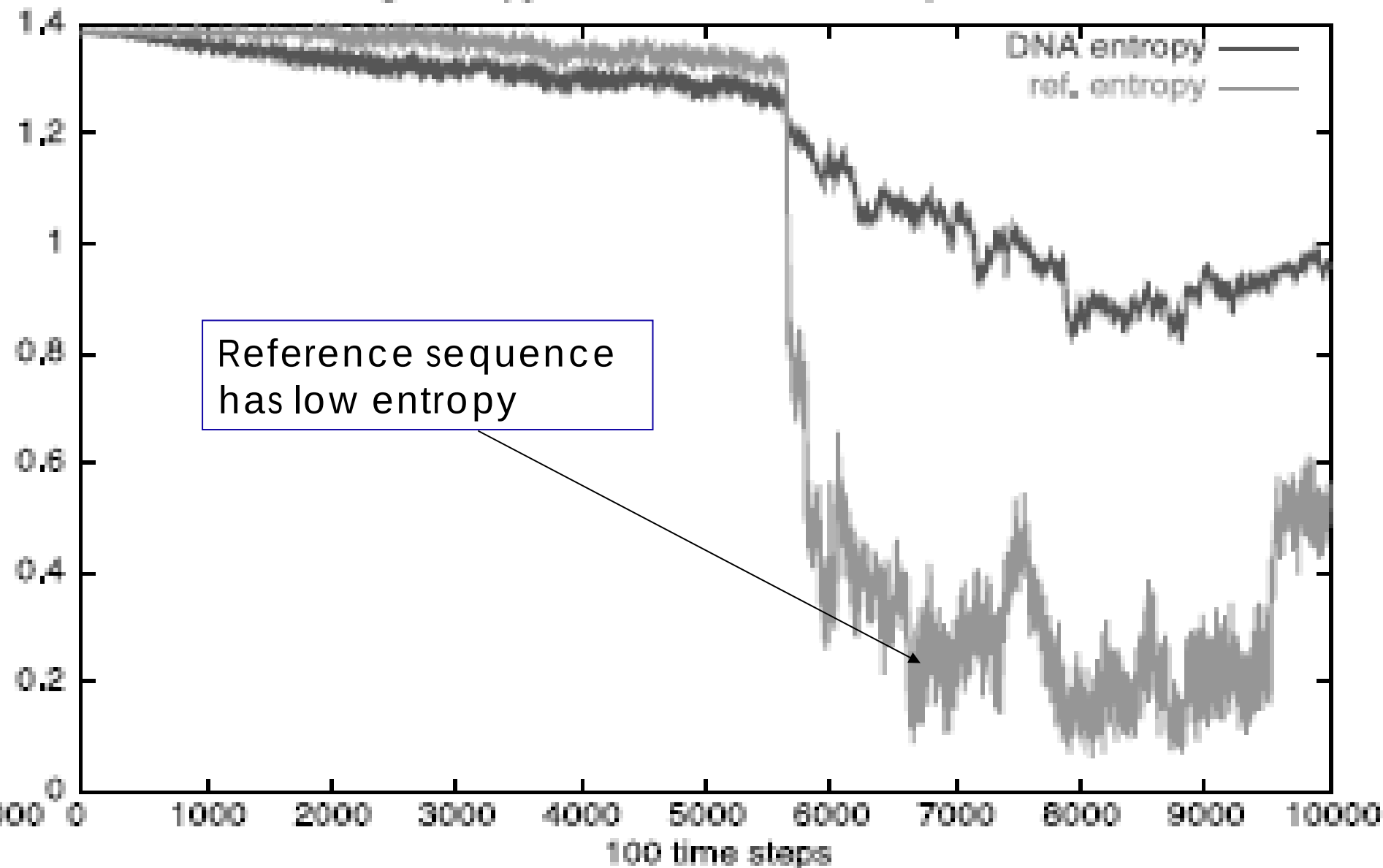
Average number of repeats of reference sequences on DNA



Cumulative number of recognized uptakes



Average entropy of DNA/reference in population



Emergence of USS depends on size of DNA

From 60 runs for each DNA length

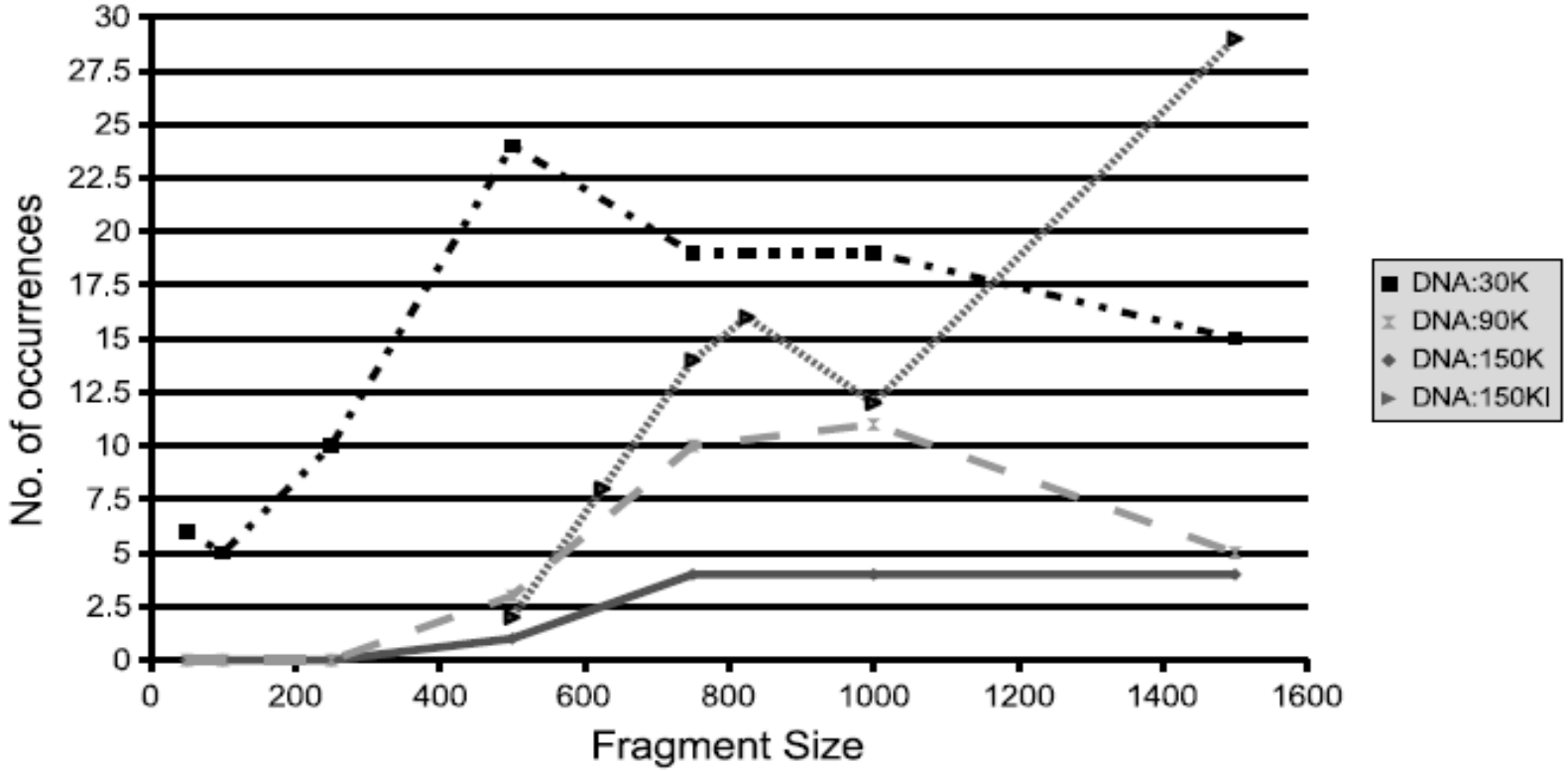


Figure 7. The number of emerged USSs in 60 runs for various DNA lengths and reference sizes. The label "DNA:30K" stands for a DNA of length 30 thousand. The 150KI series is like the 150K series, but the reference length is limited to 500 (and not to the fragment size).

Summary

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

References

- USS
  - Bakkali et al. Proc. Nat. Acad. Soc. (USA), 101 (2004) 4513-4518.
  - Chen et al. “DNA uptake signal sequences in human pathogens “, interim report. (http://sansan.phy.ncu.edu.tw/~hclee/ppr/uss_chen.pdf)
- Agent-based model:
  - http://www.brook.edu/ES/dynamics/models/history.htm
- Our papers
  - Chu et al. Artificial Life 11 (2005) 317-338.
  - Chu et al. Journal of Theoretical Biology, 238 (2006) 157-166.

People

- Biology
  - **Da-Yuan Chen**, He-Hsin Cancer Research Hospital, Taipei
  - **Rosey Redfield**, Zoology, U. British Columbia, Canada
  - **Mohamed Bakkali**, Genetics, U. Nottingham, UK
- Cellualar automata
  - **Dominique Chu/Gross**, Comp. Sci., U. Kent, UK
  - **Tom Lenaerts**, U. Libre de Bruxelles, Brussels, Belgium

THANK YOU

For papers, visit my website

(Google: HC Lee)