The Design of Genetic Codes

Chris Watkins
Royal Holloway, University of London
chrisw@cs.rhul.ac.uk

Plan of Talk

1. Raising basic questions
2. Upper bounds on quantity of information
   - Definition of channel capacity of selective breeding
   - Computation of channel capacity for genetic algorithms
   - Optimal encodings for sexual & for asexual reproduction
3. Application: possible encodings for position
4. Evolution of heritability

Basic Question 1: Quantity of Information

- Living organisms are wonderfully complex, and much information is needed to construct them
- (Most of) this information is stored in the genome
- In each generation, the information in the genomes of a species is
  - Degraded by mutations (many types)
  - Remixed by sexual recombination (if reproduction sexual)
  - Restored by selecting a fraction of the genomes produced

Basic Questions: Quantity of Information

- How much information can be maintained in the genome by evolution/selective breeding?
  - Unbounded? Or a finite limit, and if so what does it depend on?
- What encodings enable the most information to be maintained?
- How can “information from selection” even be defined?

Basic Questions (continued)

- How much information could be in the genome at mutation-selection equilibrium?
- Do the amount of information and the optimal encoding depend on whether reproduction is sexual or asexual?
- Genetic drift is random selection. How can we distinguish “directed” selection from random selection?

How much information?

- Equivalent to “How complex can organisms become through natural (or artificial) selection?”
- Amount of information in genome limits possible complexity of organism

Defining “Information from Selection”

- Many aspects of genomes may have no effect on fitness
- Small effects on fitness impractical to measure.
- Even parts of genome that *do* affect fitness may not be in fittest configuration if effect on fitness is small.
- Amount of information in the genome as a result of selection is impractical to estimate for real organisms

“Information from Selection” defined conceptually

A thought experiment:

Given a starting population of dogs, a animal breeder may breed successive generations, and select in any way he chooses.

How much information can this breeder put into the genomes of the dogs?

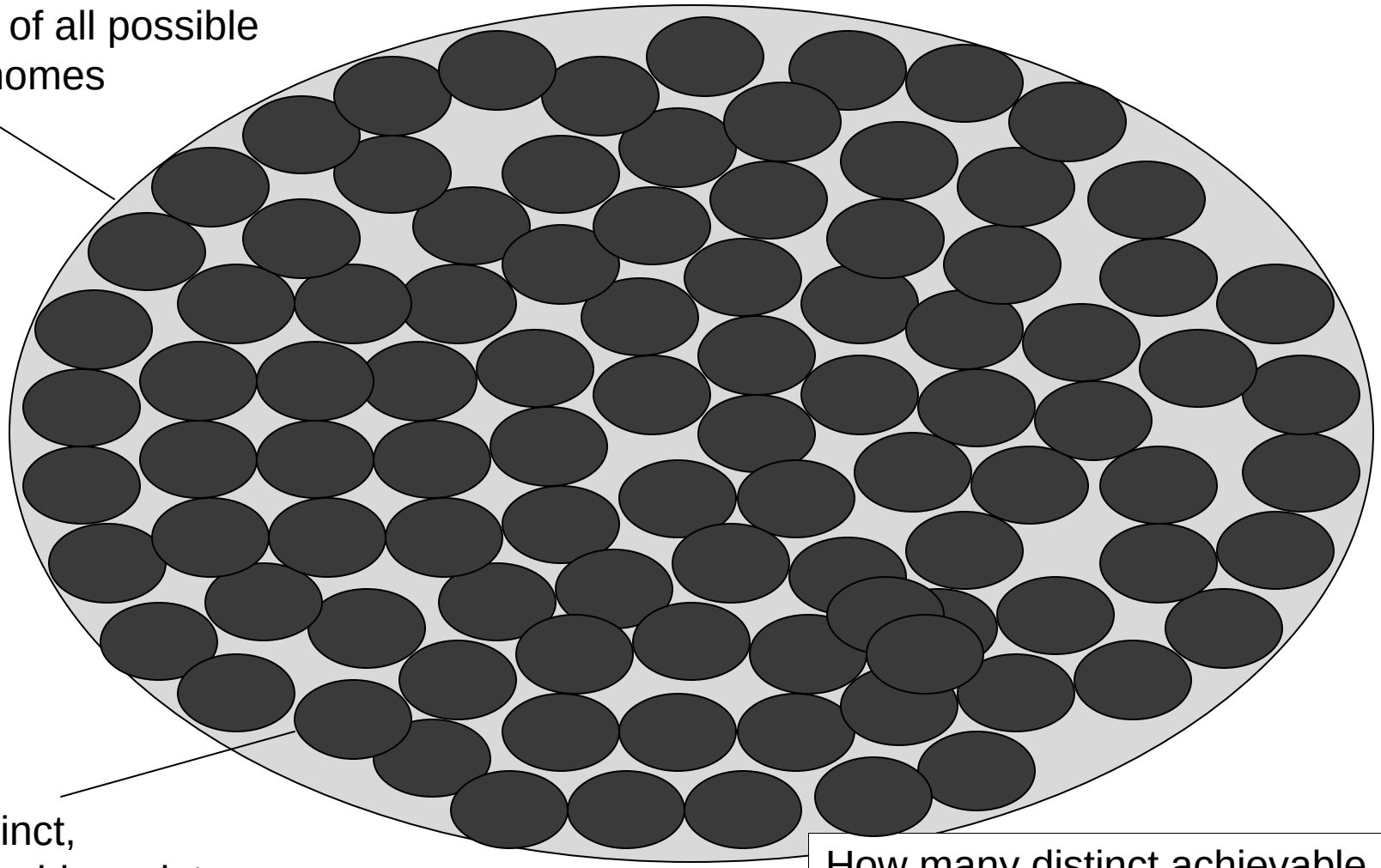
A thought experiment continued…

Suppose breeder can *reliably* create M *distinct* varieties.

Then breeder can put in at least $\log_2 M$ bits

The genome of *each* animal of a certain variety (eg a poodle) must contain information that specifies that variety.

Distinct Achievable Varieties



How many distinct achievable varieties can there be?

How many distinct achievable varieties?

- Number of distinct achievable varieties

  is less than or equal to

$$\frac{\textit{Number of all possible genomes}}{\textit{No. of genomes in an achievable variety}}$$

Channel Capacity

A cleaner formalisation of the same intuition:

Consider a “communication channel” in which the sender is the breeder, and the receiver is a naturalist who receives a single animal from the breeder’s final population.

Channel Capacity (continued)

The breeder and the naturalist may confer beforehand, and share all information (starting population, breeding setup, etc) ***except*** the selection policy the breeder will use.

They may agree any coding system they choose.

Channel Capacity (continued)

The ***message sent*** S is the selection policy that the breeder follows.

The ***message received*** R is the genome of a single organism from the breeder's final population.

Channel Capacity (continued)

The ***channel capacity*** is the maximal achievable amount of information that can be sent through this “communication channel”.

This is a natural “single figure” measure of the extent to which the breeder can influence the genomes that are produced.

Channel Capacity (continued)

S is generated according to sending distribution Q

Information sent is:

$I(S;R) = H(S) - H(S|R) = \mathbf{H(R) - H(R|S)}$

Channel capacity found by maximising I(S;R) over all possible Q

Selective Breeding as Communication

- A geneticist Alice is imprisoned; she wishes to send a message to Bob who is outside
- Only way: Alice captures *Drosophila* and selectively breeds them in her cell
- Alice's message encoded in the genomes of her flies, by selective breeding only.

Selective Breeding as Communication

- On the previously agreed day, Alice releases her final population: Bob captures **one** fly, and decodes message
- Alice and Bob have previously agreed a code.
- How much information can Alice send to Bob?

Selective Breeding as Communication

- What code should Alice and Bob use?
  - small no. of loci, well controlled?
  - large no. of loci, poorly controlled?
  - sexual or asexual organism?

Channel Capacity

- Channel capacity is maximal amount of information Alice can expect to send under the most optimistic assumptions
  using the best coding system

- Depends on number of generations, population size, mutation rate, selection intensity

Channel Capacity of Evolution

- Channel capacity is a measure of maximal **evolvability**
- Lineages with greater channel capacity have potentially greater range of adaptive response

Genetic Algorithms

- Very simple abstraction of genetics
- Genomes are haploid binary sequences of fixed length L
- Mutation rate U – probability of inverting a bit, independently of other bits
- No insertions/deletions
- Breeding either asexual (mutation only), or sexual with each bit independently chosen from either parent (“uniform crossover”)

Channel Capacity of GA

Only need to use fitness functions of the form:

F(g) = fraction of agreement with g*, for some chosen ideal genome g*

Selection policy: select “fittest” 50% at each generation (truncation)

2^L possible selection policies

Choose sending distribution Q as uniform distribution over possible choices of g*

H(R) = L , by symmetry

Channel Capacity of GA (cont.)

What is H(R|S) ? Possible distribution of genomes in equilibrium state of GA is complicated…

If p is the mean fraction of agreements with g* at equilibrium, can use bound from maximum entropy distribution with expected fraction p of agreements.

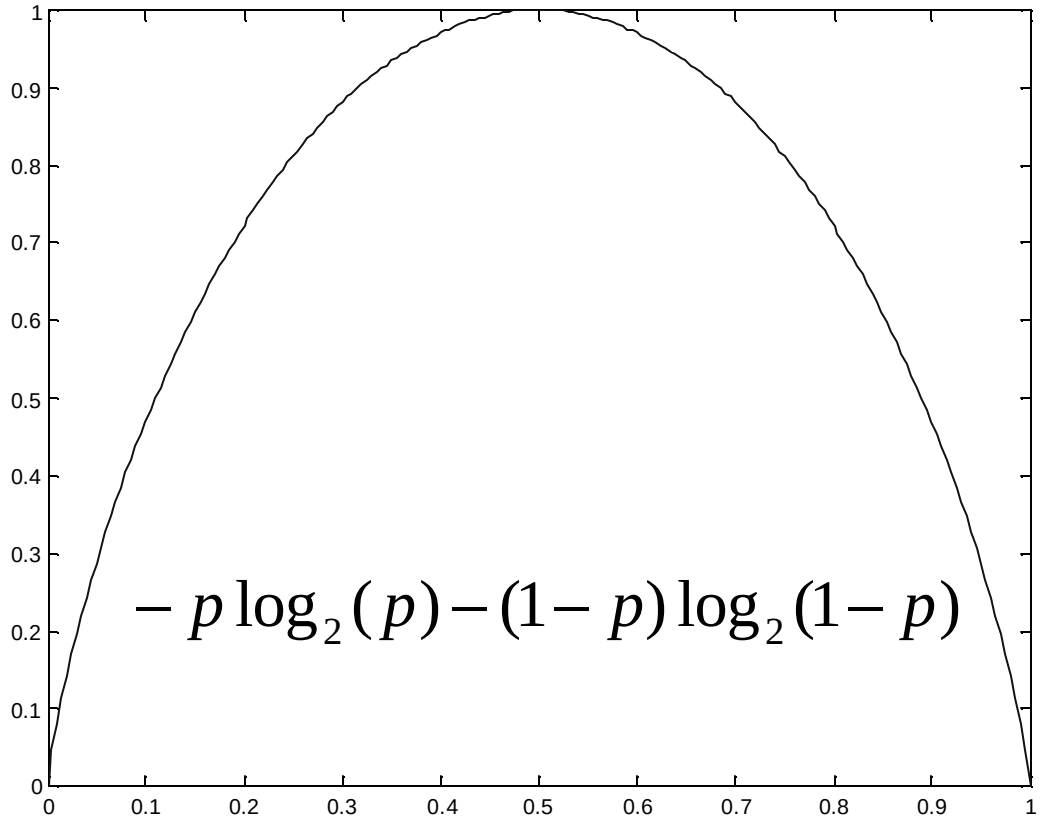
Max ent. Distn. is factorial distribution, with probability of agreement of p at each locus independently – entropy is L h(p), where h is entropy of a Bernoulli variable with parameter p.

H(R|S) <= L h(p)

Entropy: h(p)

Information

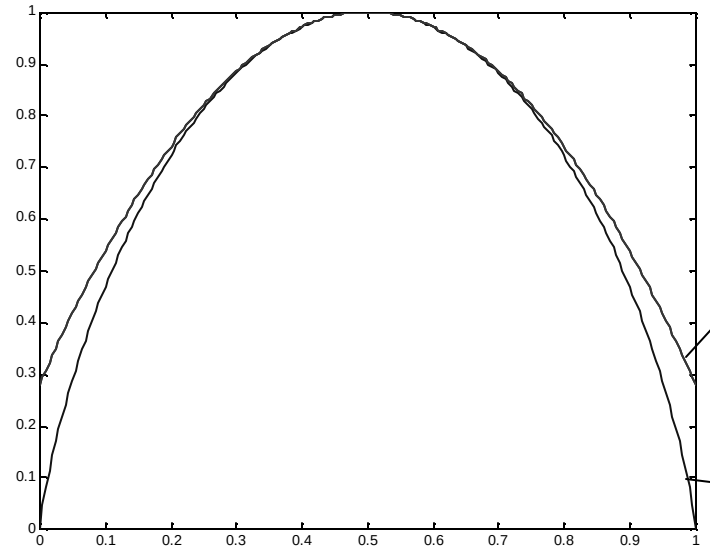
(bits)



p

Quadratic approximation of h(p) near p = ½

Second order Taylor expansion about p = ½



p

(probability)

$$1 - \frac{2}{\ln 2}\left(p - \frac{1}{2}\right)^2$$

$$-p\log_2(p) - (1-p)\log_2(1-p)$$

Channel capacity where p close to ½

$$C \approx \frac{2}{\ln 2} L \left(p - \tfrac{1}{2} \right)^2$$

C will approach a limit for large L if

$$p - \frac{1}{2} = O\left(1 / \sqrt{L} \right)$$

What is p at large-population equilibrium?

w.l.o.g. suppose g* is 11111111….111

At mutation-selection equilibrium with mutation rate U:

Fraction of zeros introduced by mutation is:

$$2\,U\,(p - \tfrac{1}{2})$$

Fraction of zeros eliminated is (approx) equal to standard deviation of fractions of zeros in individual genomes. We can bound this from above as:

$$\sqrt{\frac{p(1-p)}{L}}$$

What is p at equilibrium?

Solving

$$2U\left(p-\tfrac{1}{2}\right)=\sqrt{\frac{p(1-p)}{L}}$$

We obtain:

$$p=\tfrac{1}{2}+\frac{1}{2\sqrt{1+4LU^{2}}}\approx\tfrac{1}{2}+\frac{1}{4U\sqrt{L}}$$

Channel Capacity at Equilibrium

- For sexual reproduction,
  large population size N (> 1/U)
  large genome size L,
  p close to ½:

$$ChannelCapacity \propto L(p-\tfrac{1}{2})^2 \propto \frac{1}{U^2}$$

Asexual Reproduction

- Simplest to consider **strong** selection in large-genome limit.
- Population of N individuals.
- In each generation, select single best individual, and breed N children from this individual
- Key point: strong selection is **most effective possible** form of selection for asexual reproduction with limit on population size

Asexual Reproduction with Strong Selection

For large L, small U, mean and variance of number of new mutations per individual is LU.

At equilibrium, expected fitness of best child is equal to fitness of parent.

Fitness (fraction of ones) of children distributed approx normally (for large L) with variance U/L

Channel Capacity at Equilibrium

- For asexual reproduction,
  population size N,
  large genome size L,
  strong selection:

$$ChannelCapacity = \frac{O(\log N)}{U}$$

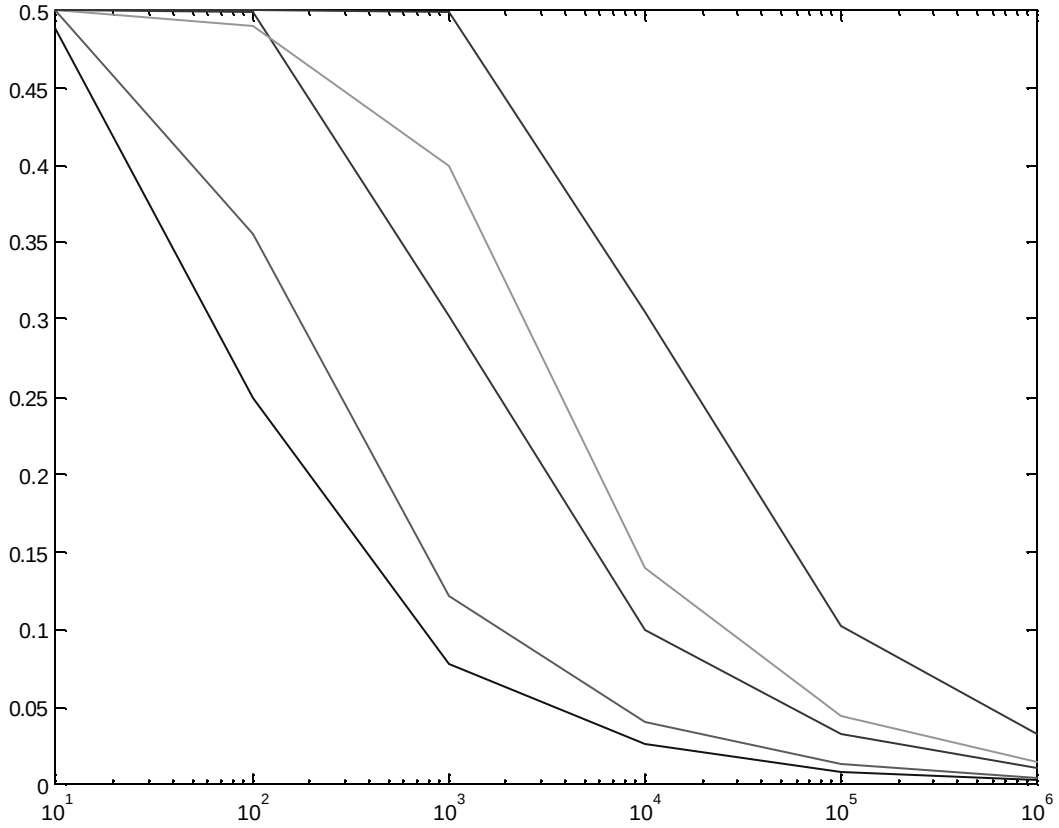
Channel Capacity at Equilibrium

- For asexual reproduction, large genome size L, **truncation** selection:

$$ChannelCapacity = O\left(\frac{1}{U}\right)$$

Asexual: (P-0.5) vs L

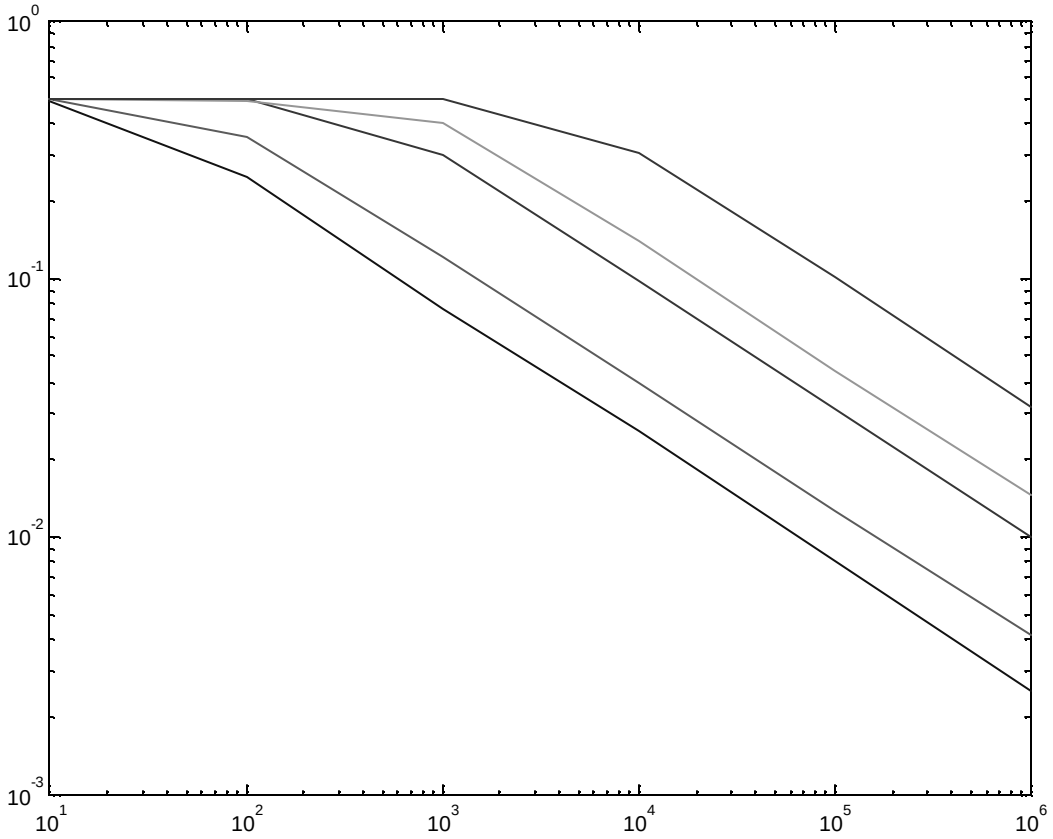
Fraction of “good” alleles minus 1/2



Genome length L (log scale)

Asexual:(P-0.5) vs L

Log of Fraction of “good” alleles minus 1/2



Genome length L (log scale)

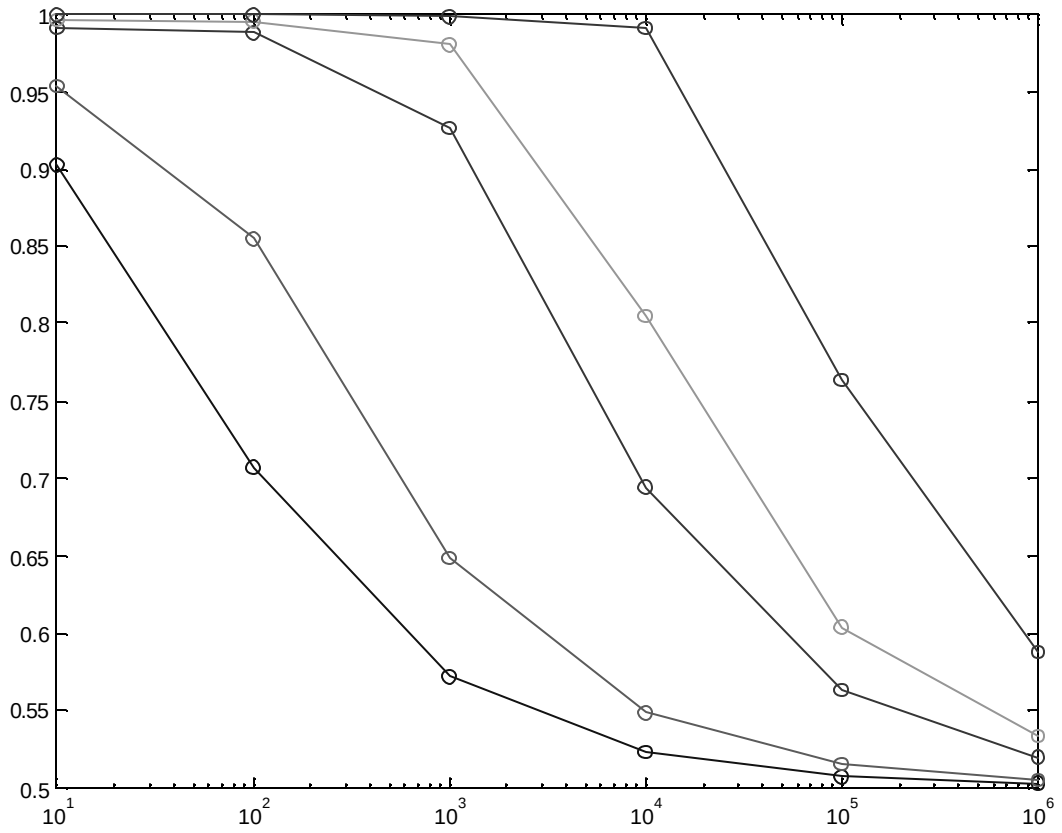
Sexual: p vs L

Population: 500 Multiplicative Selection

1

Fraction of “good” alleles minus 1/2

P

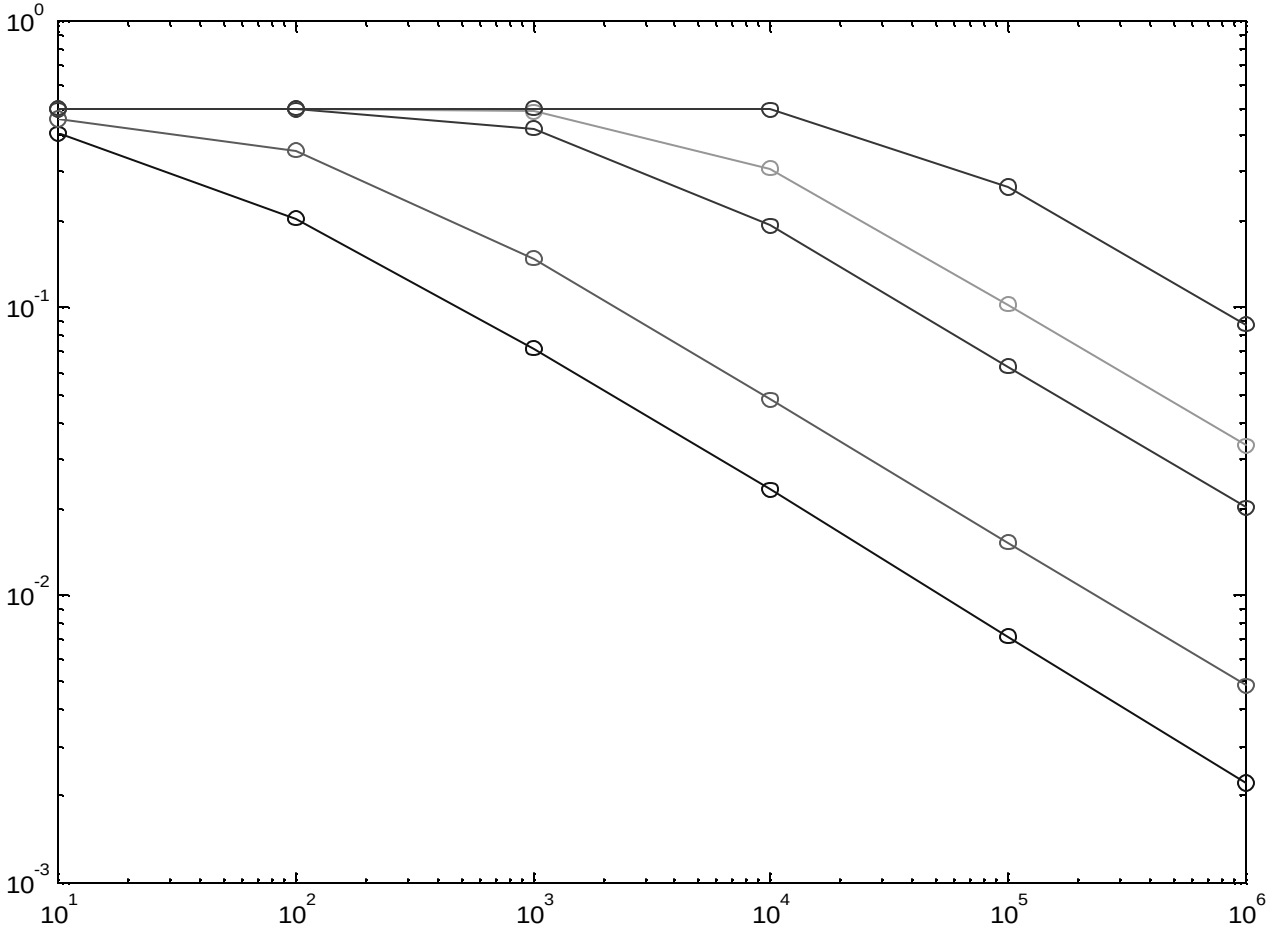


½

Genome length L (log scale)

Sexual: log(p-1/2) vs L

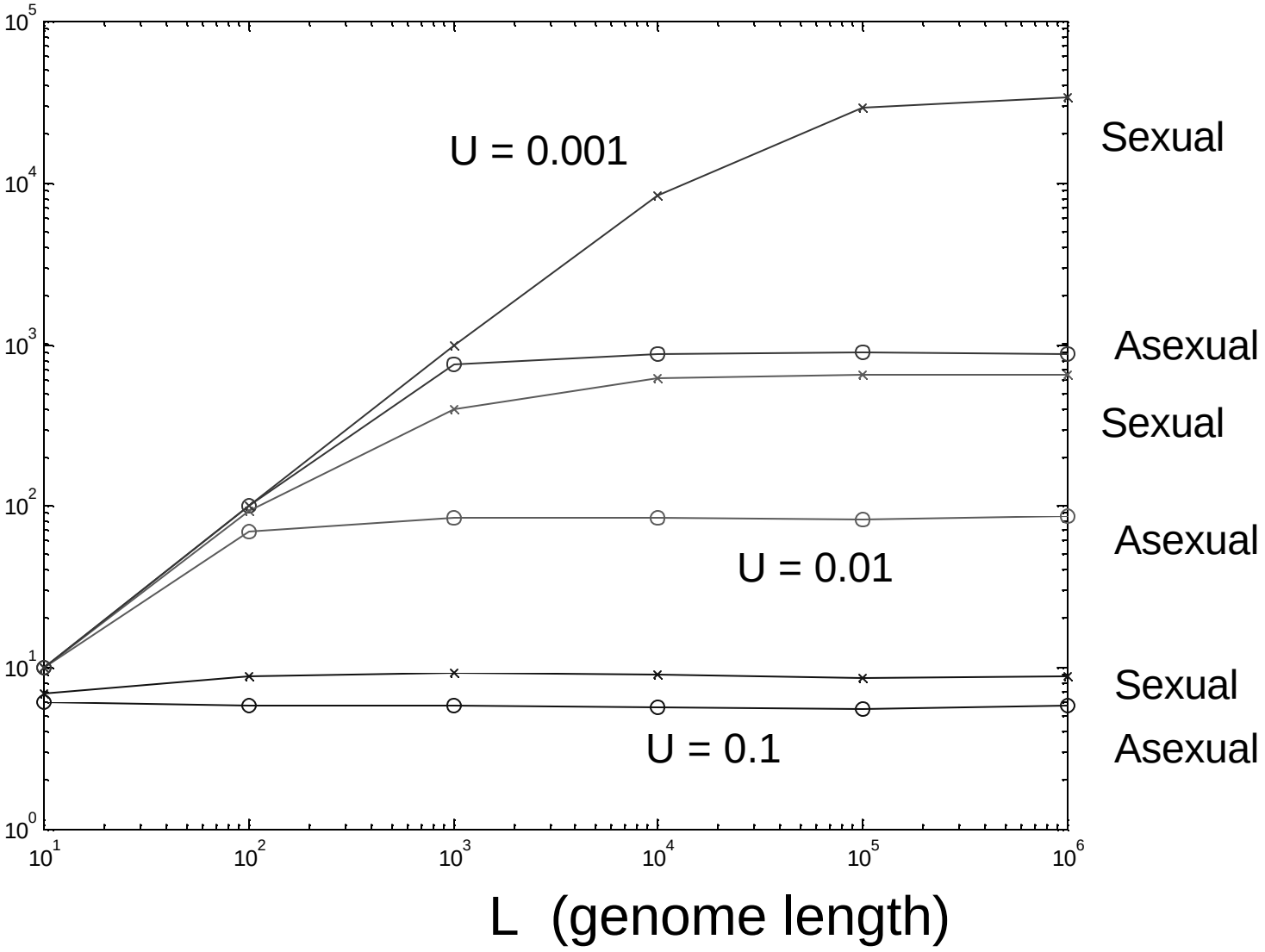
Log of Fraction of “good” alleles minus 1/2



Genome length L (log scale)

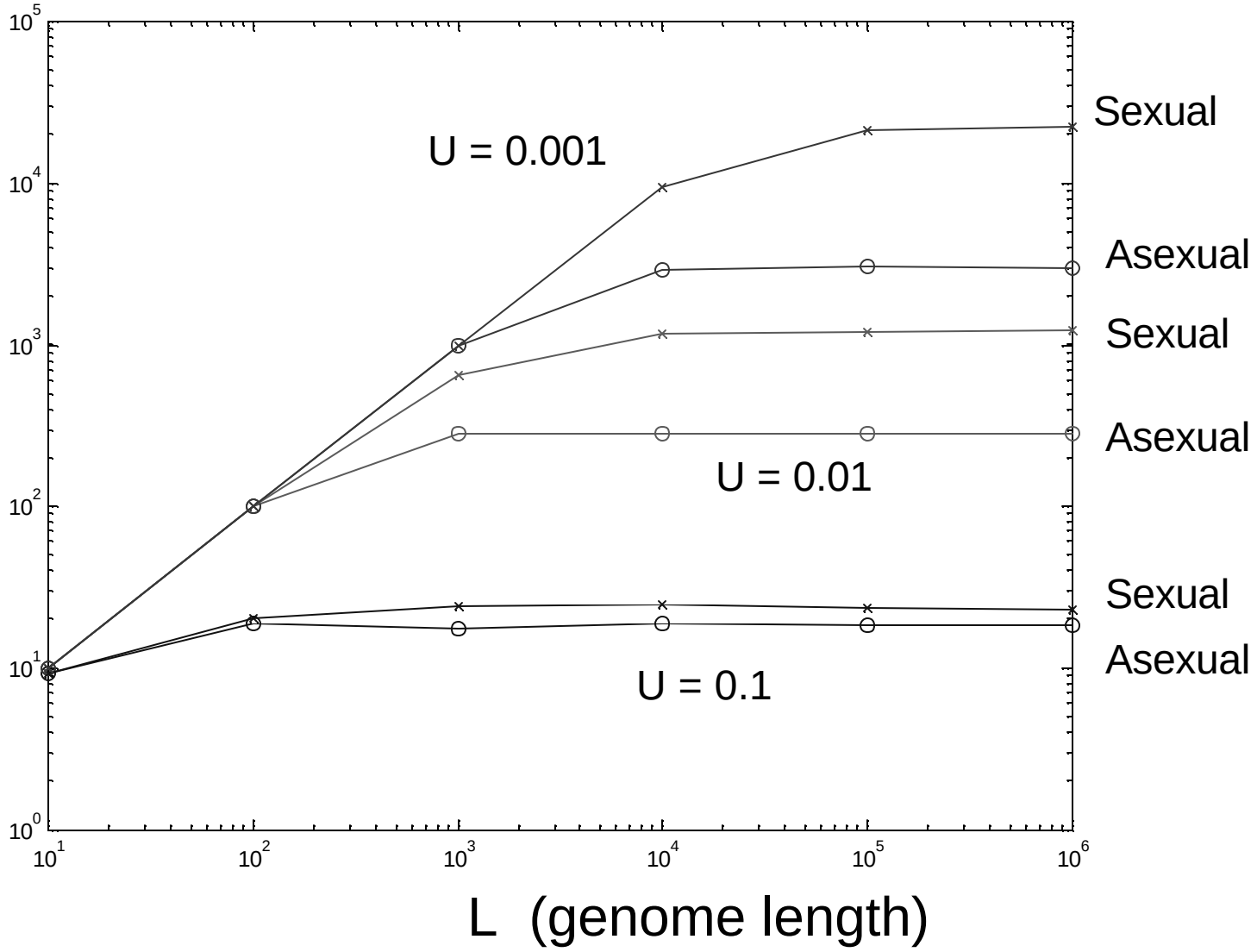
Equilibrium Information vs Genome Length
Truncation Selection (50%)

Population = 500



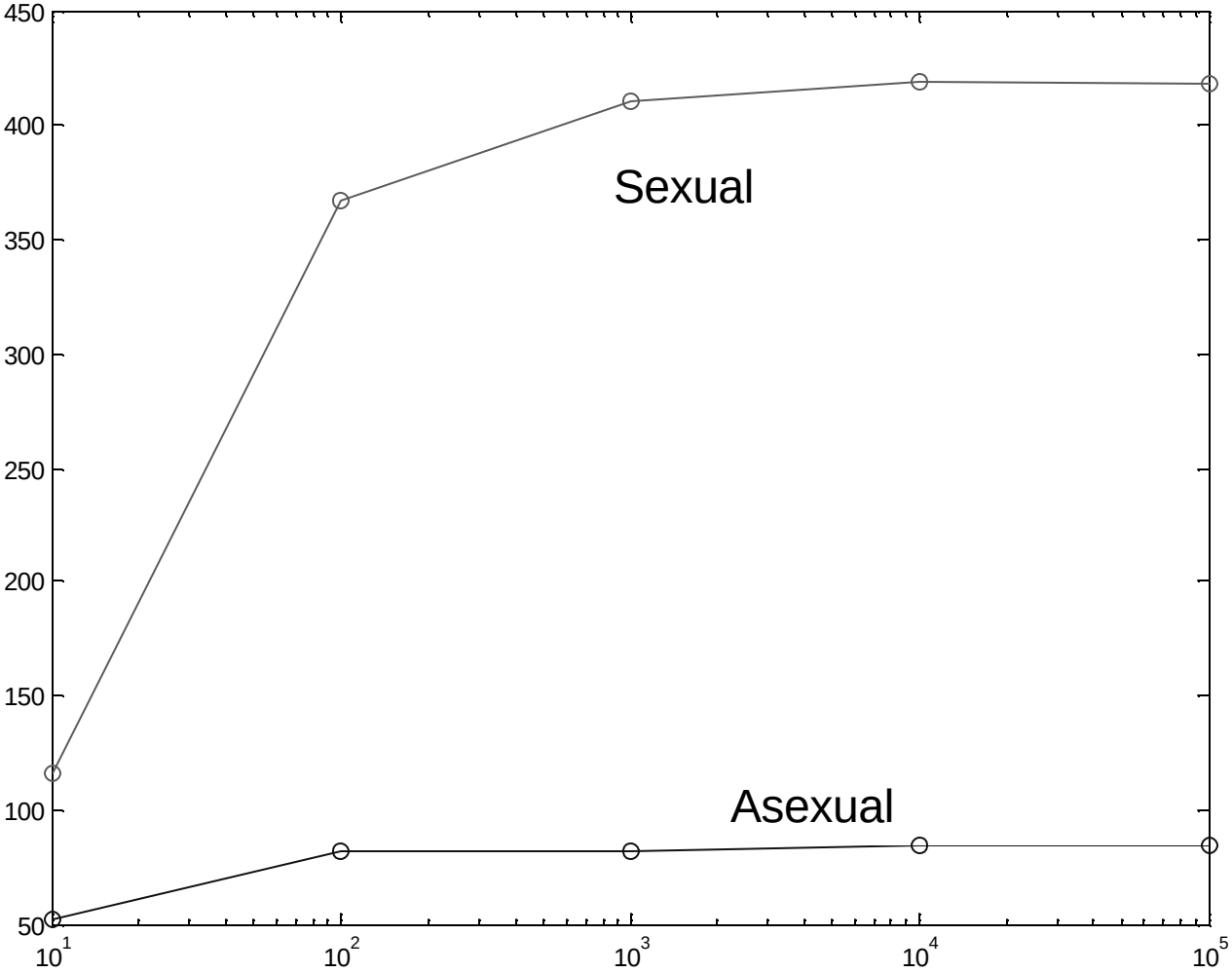
Equilibrium Information vs Genome Length
Multiplicative Selection (Intensity equivalent to 50% truncation)

Population = 500



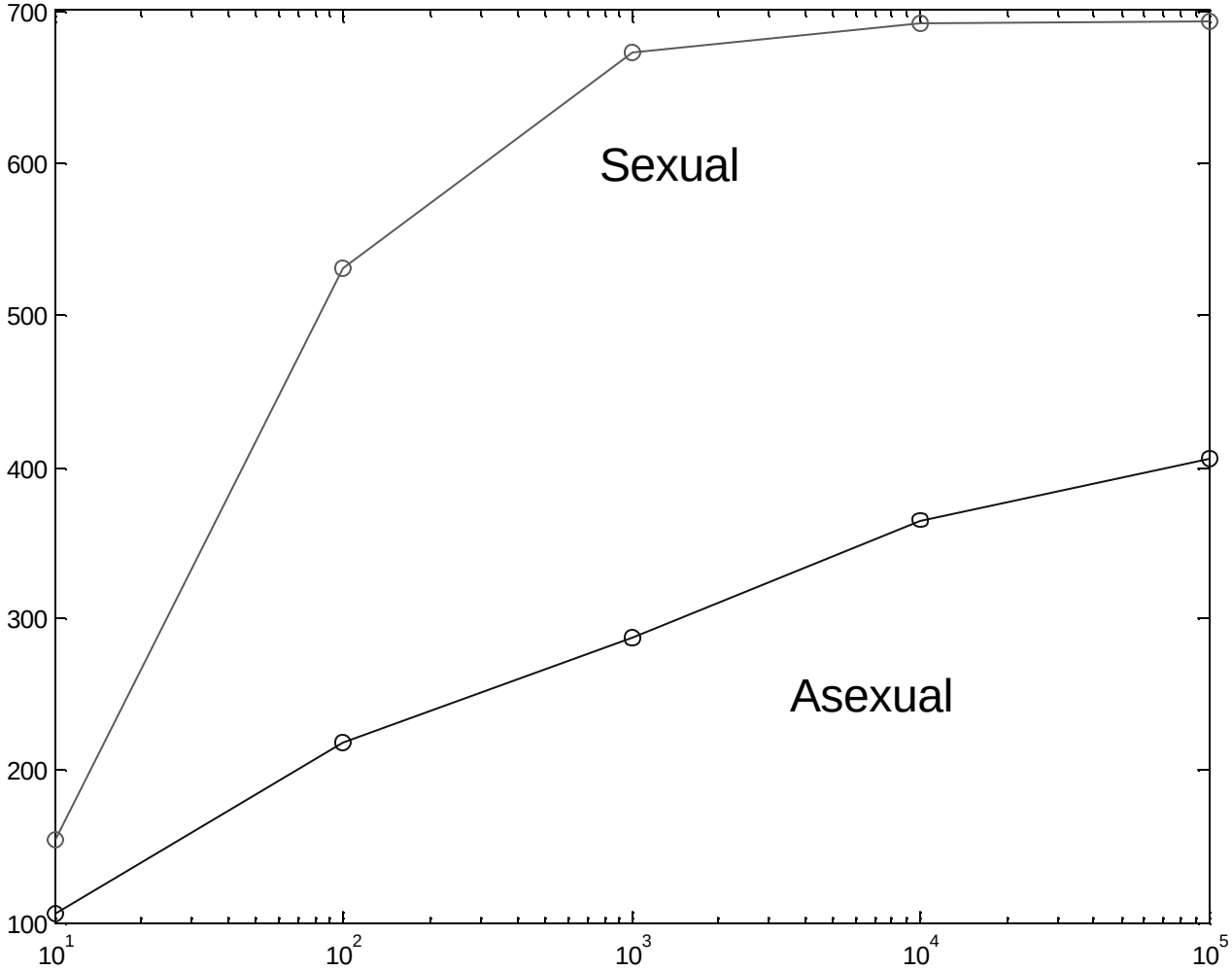
Info vs Pop size (truncation)

L = 1000 U = 0.01 KLD = ln 2



Info vs Pop Size (Multiplicative selection)

L = 1000 U = 0.01 KLD = ln 2



Discussion

- GAs have high channel capacity with sexual reproduction and LARGE genomes

- Channel capacity of sexual reproduction is higher than that for asexual reproduction. For low mutation rates, difference is enormous.

Discussion

- Informationally optimal genetic encoding for sexual reproduction is:
  - dispersed (spread out)
  - distributed (many loci)
- Many differences between eukaryotic and prokaryotic genomes make sense from this point of view

Two kinds of Cell

Prokaryotes (bacteria) – small, single-celled, asexual, no nucleus, simple structure, one circular chromosome of DNA

Eukaryotes – larger cells, many chromosomes of DNA in nucleus, complex structure, often multi-celled organisms, sexual reproduction

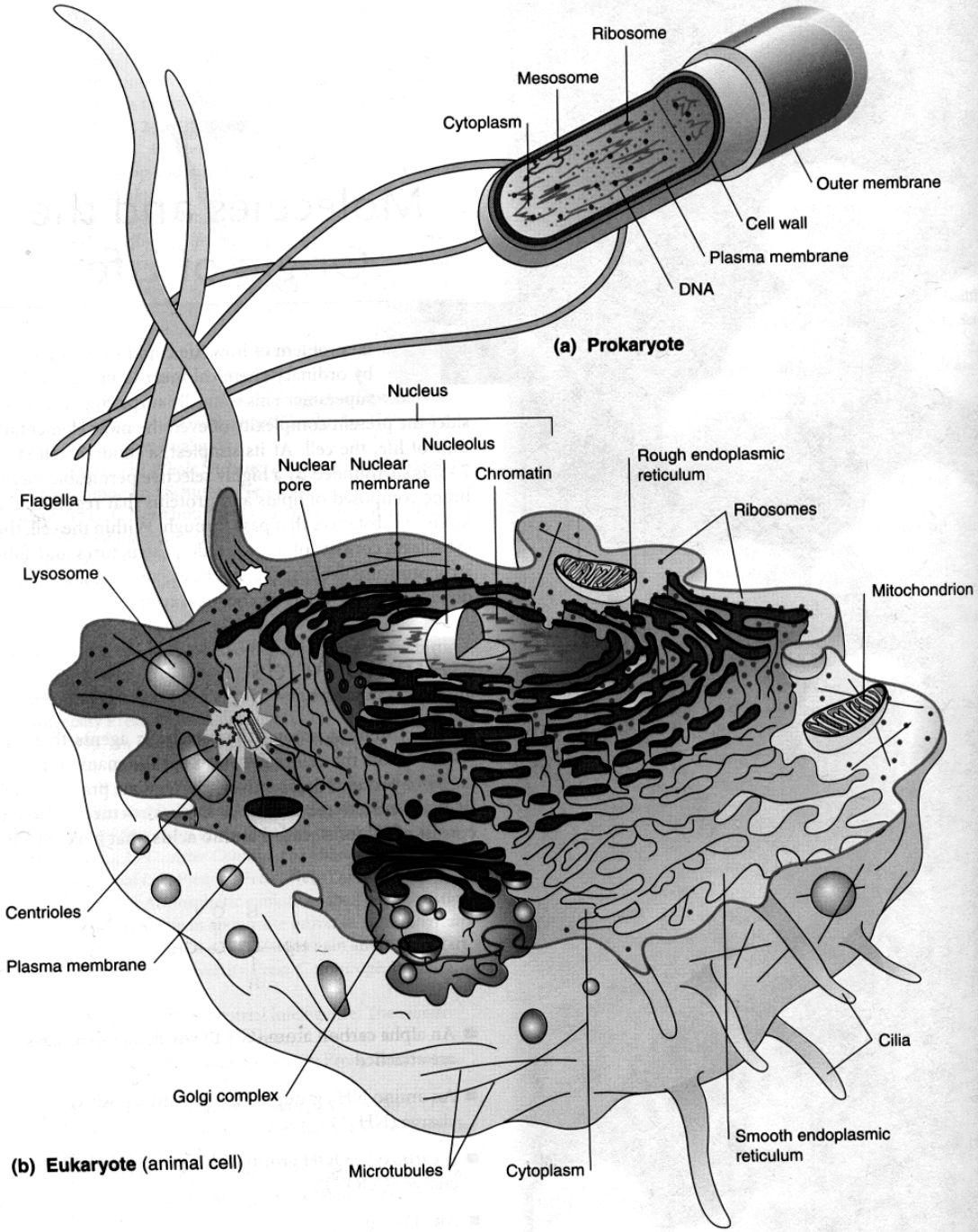


Figure scanned from “Evolution” by Monroe W. Strickberger, Jones & Bartlett Publishers, 2000

How much DNA?

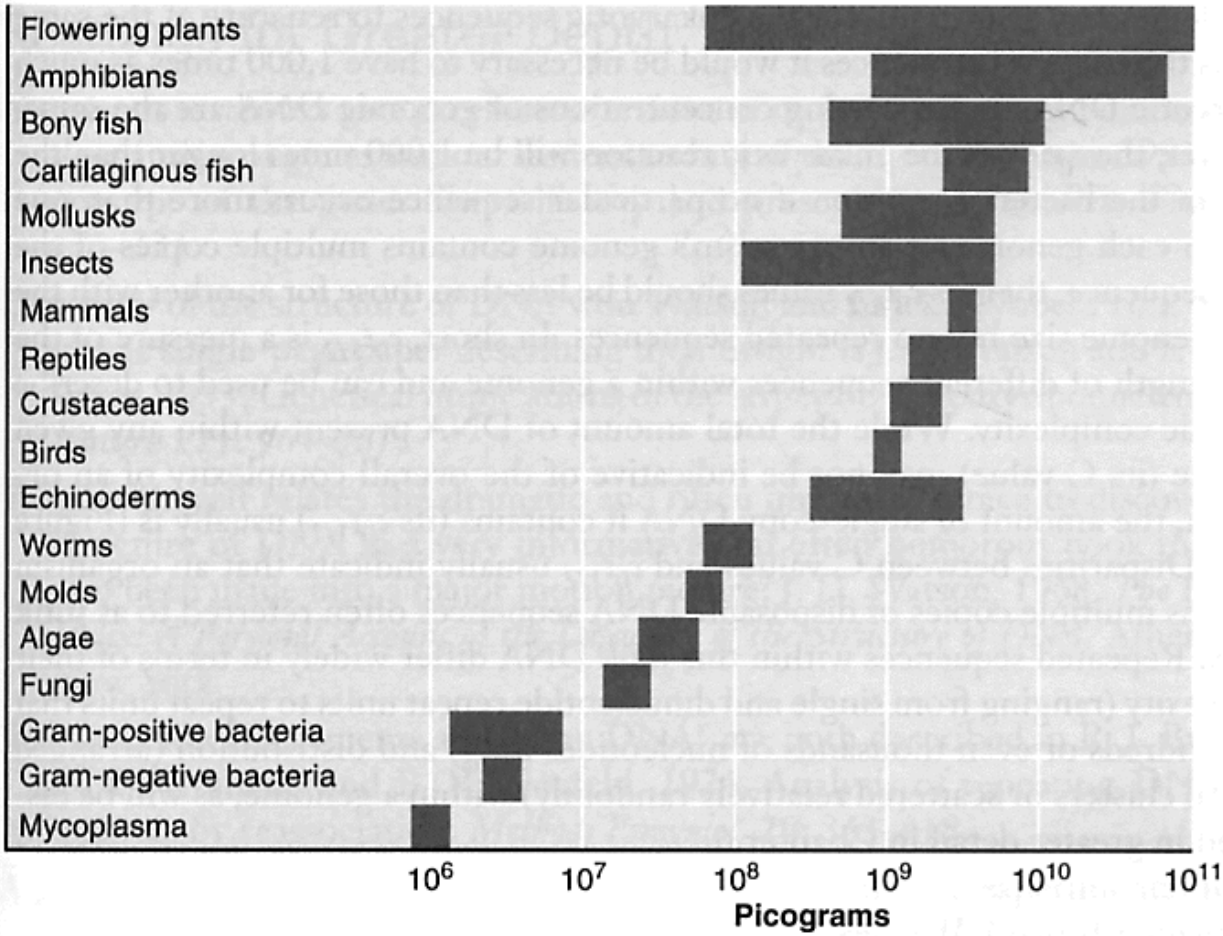


FIGURE 1.14 *The DNA contents of the haploid genomes of a variety of different organisms. C values are generally correlated to morphological complexity in simpler eukaryotes but vary significantly among more complex eukaryotes. The range of DNA content within a phylum is indicated by the shaded areas.*

Minimum genome sizes in various classes of organism

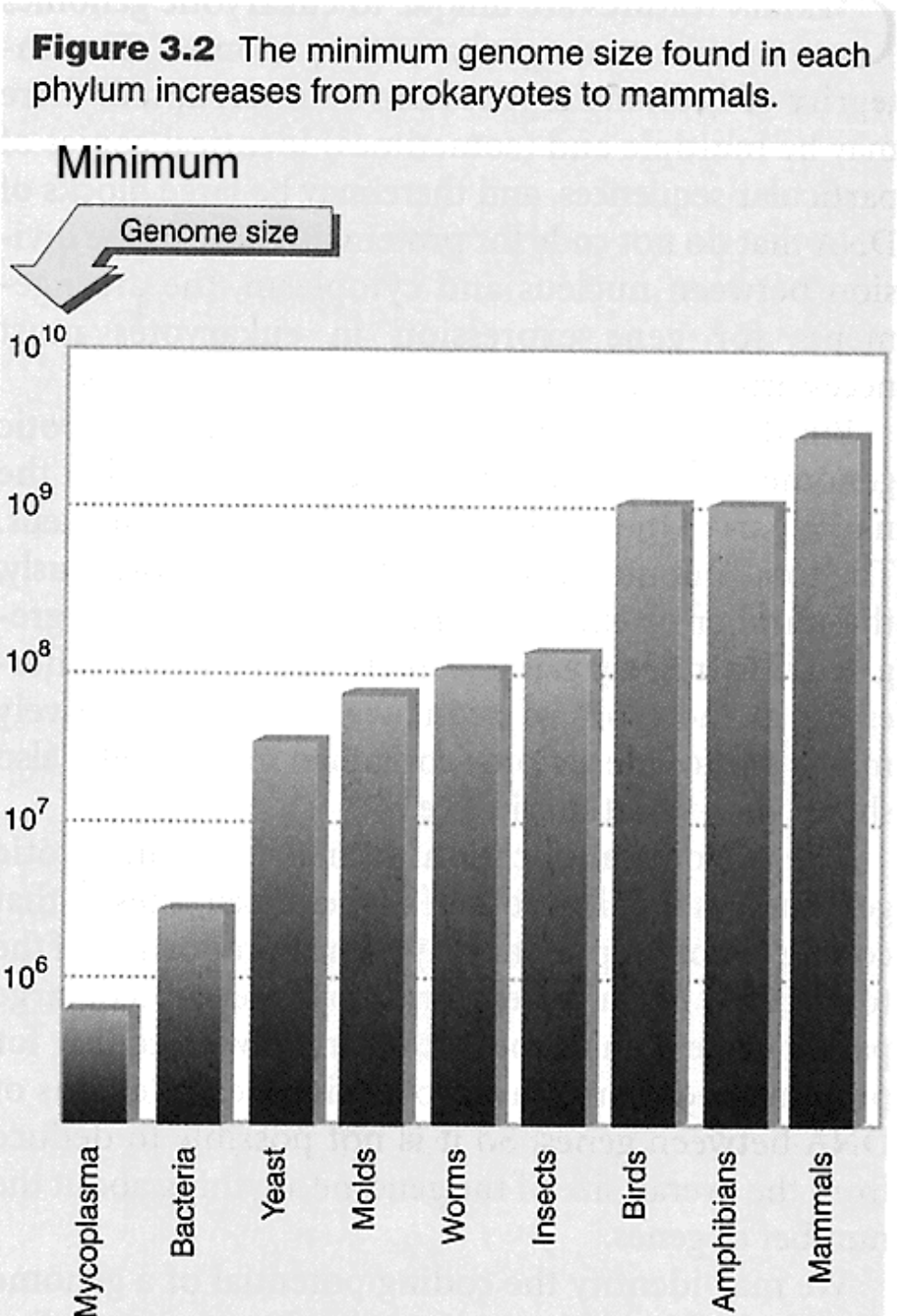


Figure 3.2 The minimum genome size found in each phylum increases from prokaryotes to mammals.

Figure scanned from “Genes VII” by Benjamin Lewin, OUP 200

Speculations

- A compact code such as the triplet genetic code is efficient for asexuals (prokaryotes)
- Many aspects of eukaryotic genetic encoding can be viewed as informationally efficient distributed and dispersed encodings:
  - Introns and alternative splicing
    - Dispersed splicing signals?
    - Why is gene-finding difficult??
  - Larger proteins
  - More metameric proteins
  - Gene regulation by many weak enhancers far from gene
  - More post-translational modification