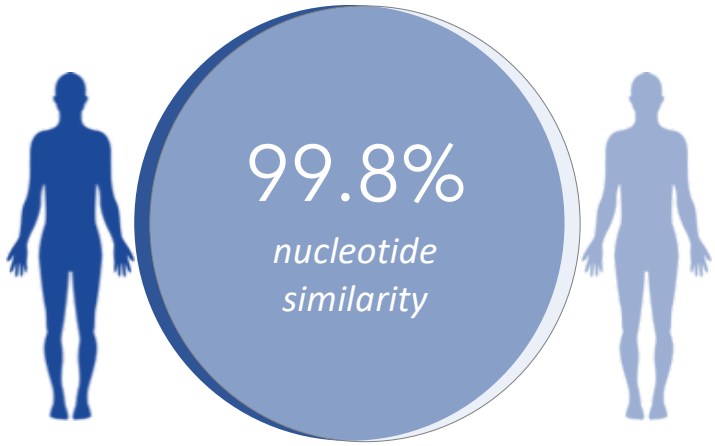
The neutral and nearly neutral theories of molecular evolution

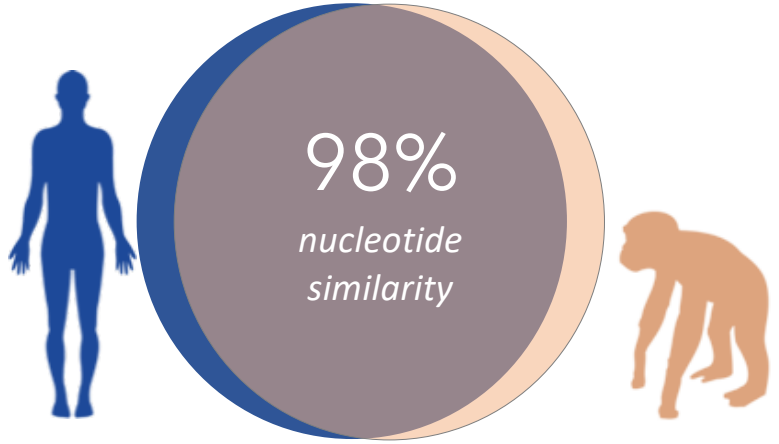
Joseph P. Bielawski

Department of Biology
Department of Mathematic and Statistics
Institute of Comparative genomics
Dalhousie University, Halifax, Nova Scotia, Canada

Models of molecular evolution seek to explain the *origin* and *maintenance* of genetic variation



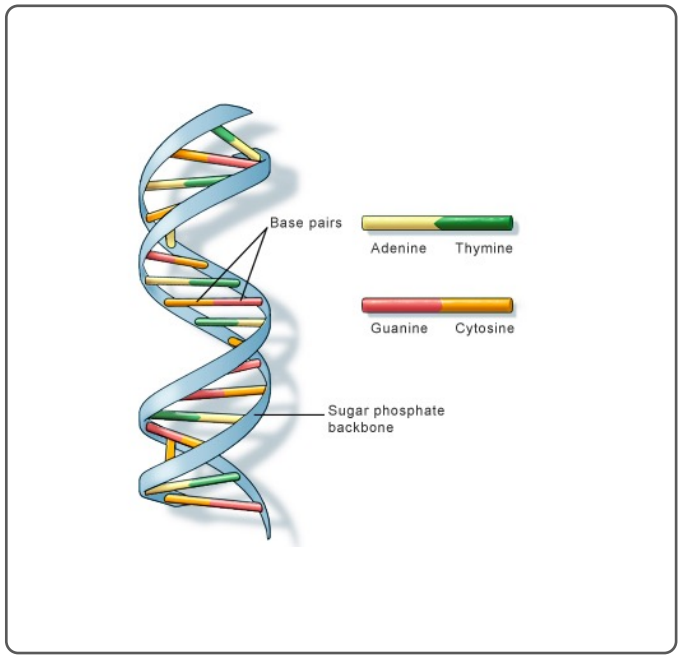
~ 5 million nucleotide differences



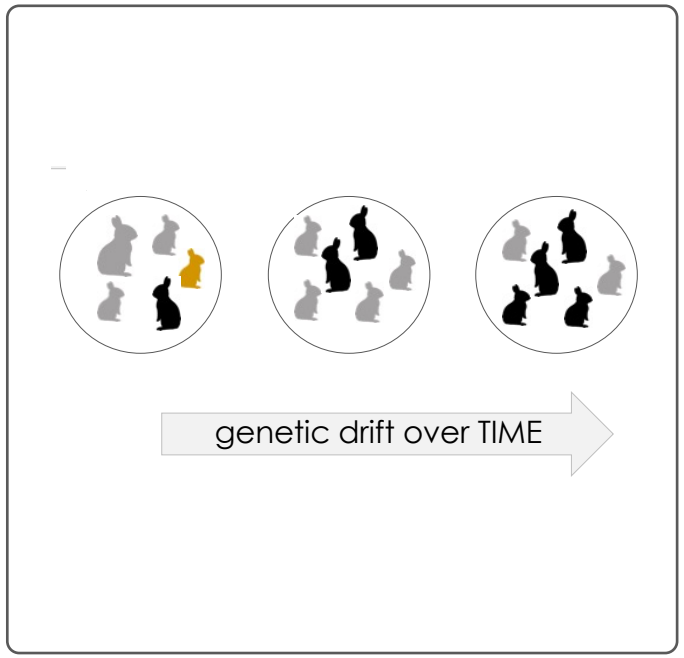
~ 35 million nucleotide differences

brief review of the fundamental evolutionary “forces”…

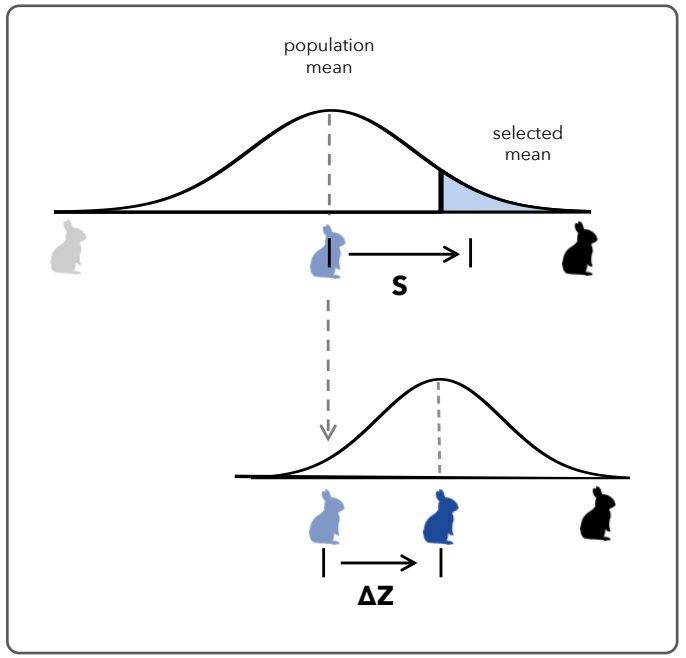
1. mutation



2. drift

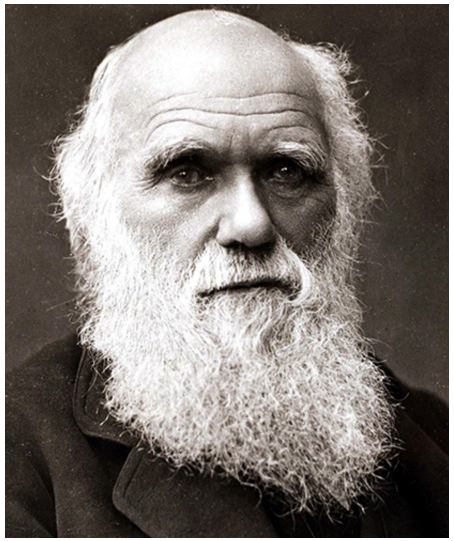


3. natural selection



Conceptual models that explain genetic variation:

1. Neo-Darwinism



natural selection dominates

2. Neutral Theory

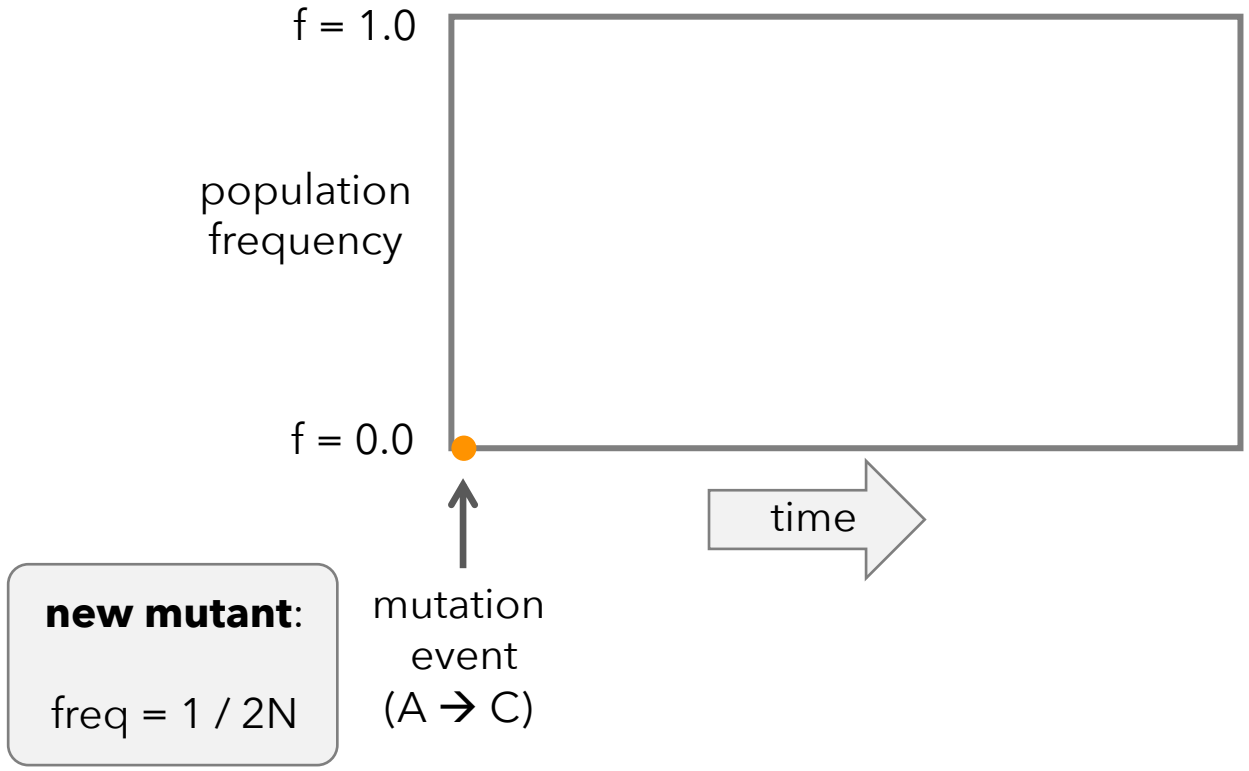
genetic drift dominates

3. Nearly Neutral

drift and selection **interact**

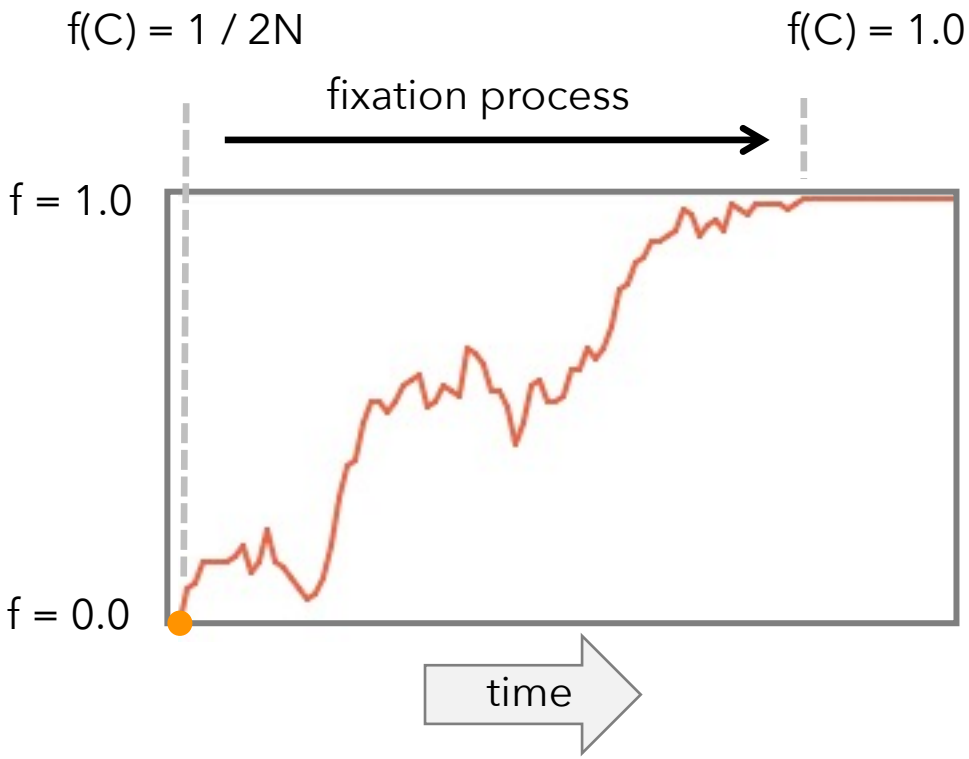
Key population concepts:

1. **mutation**
2. fixation
3. substitution



Key concepts:

1. mutation
2. **fixation**
3. substitution



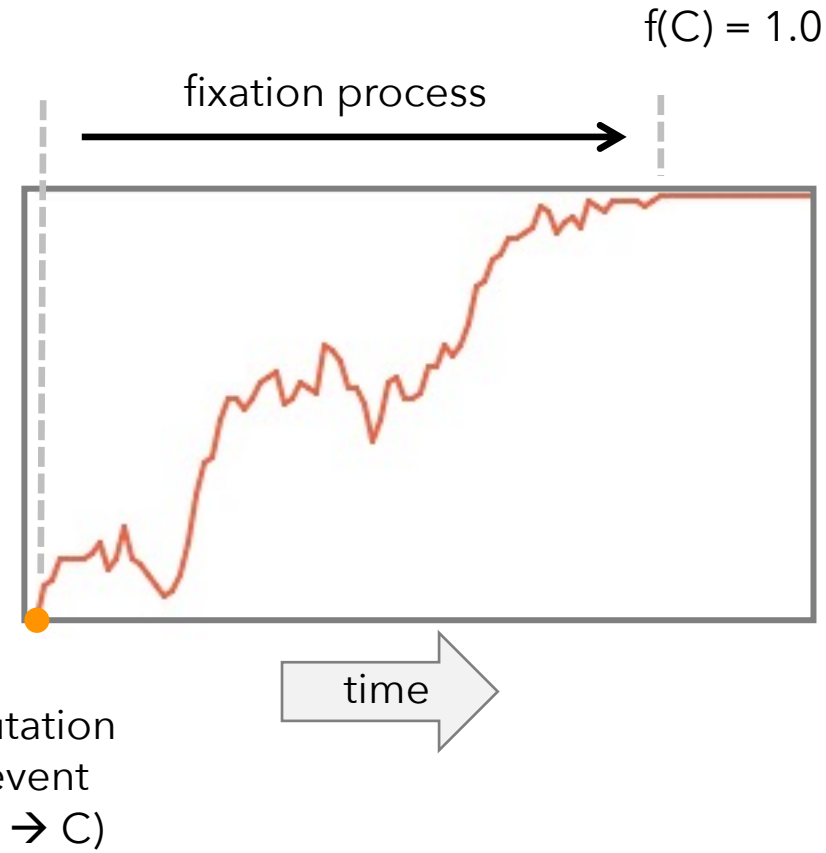
fixation:
by drift process **or**
selection process

Key concepts:

1. mutation
2. fixation
3. **substitution**

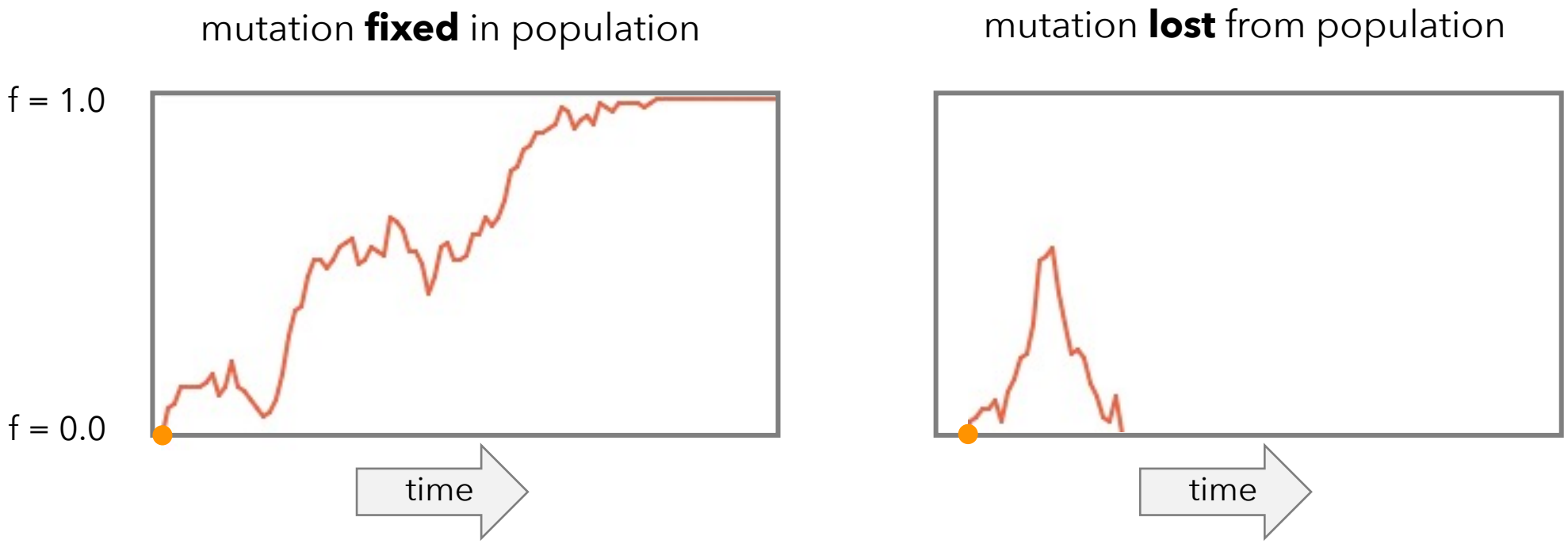
Macro-evolutionary models:
… often "***screen off***" micro-evolution

substitution:
the outcome of fixation process



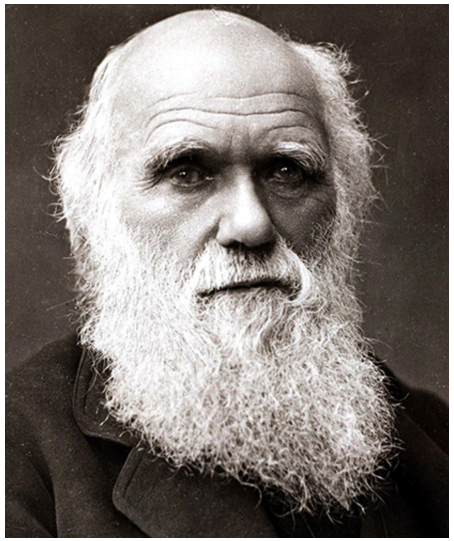
substitution:
change in "state" of the population (A → C)

Mutations can be **fixed** or **lost**



Conceptual models that explain genetic variation:

1. Neo-Darwinism



natural selection dominates

2. Neutral Theory

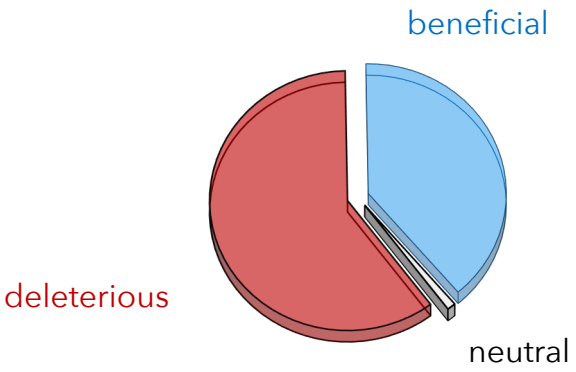
genetic drift dominates

3. Nearly Neutral

drift and selection **interact**

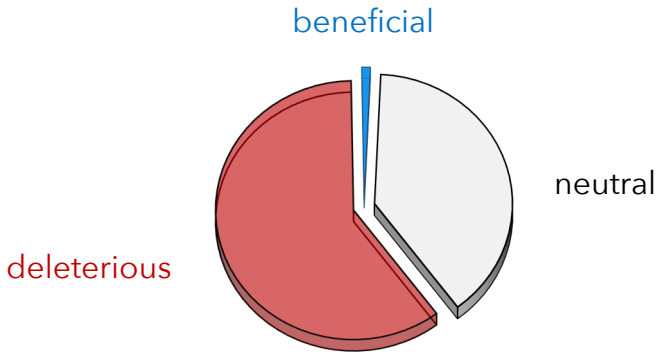
Conceptual models that explain genetic variation:

1. Neo-Darwinism



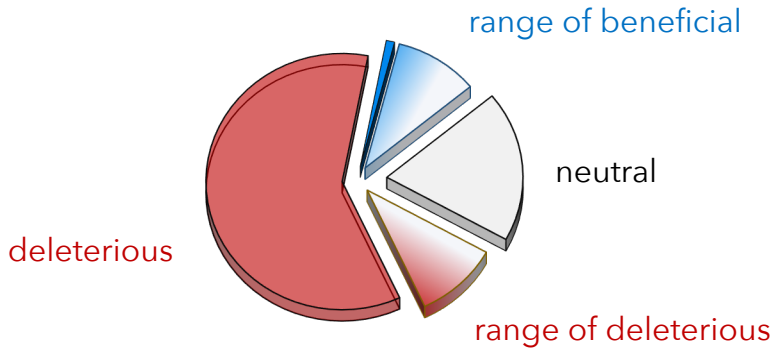
natural selection dominates

2. Neutral Theory



genetic drift dominates

3. Nearly Neutral



drift and selection **interact**

Conceptual models that explain genetic variation:

1. Neo-Darwinism

natural selection dominates

2. Neutral Theory

genetic drift dominates

3. Nearly Neutral

drift and selection **interact**

Neutral theory of molecular evolution (Kimura 1968)

k = rate of nucleotide substitution [**at mutation-drift equilibrium**]

k = new mutations × probability of fixation

the **number of new mutations** arising in a diploid population

$$2N\mu$$

the **fixation probability** of a new mutant by drift

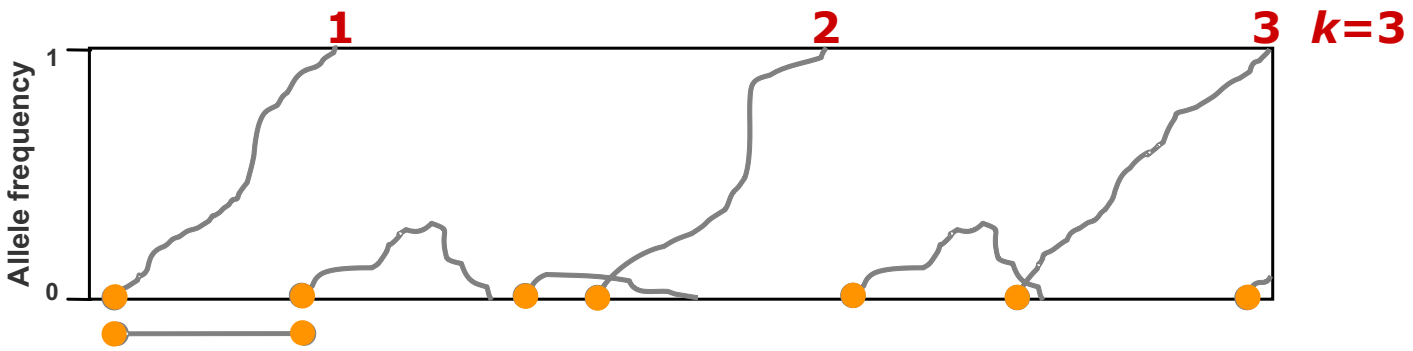
$$1/2N$$

the **substitution (fixation) rate**, *k*

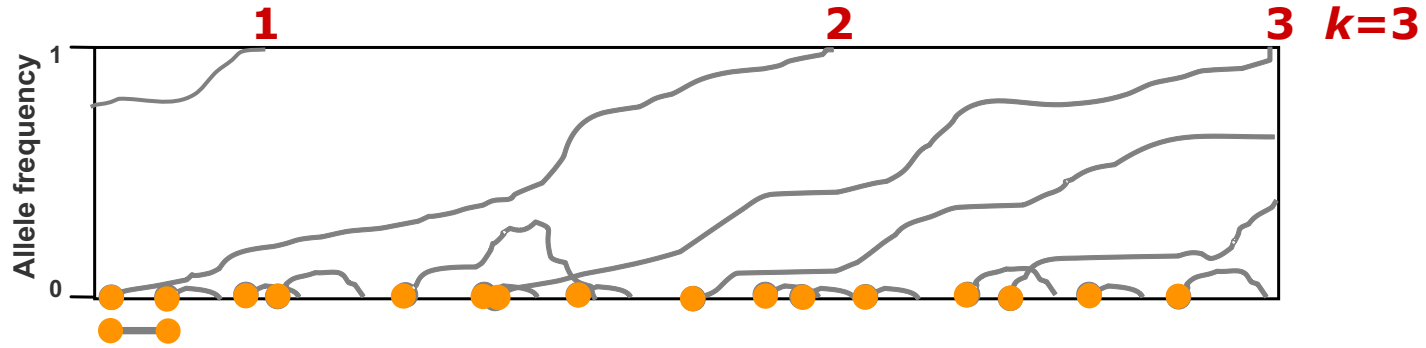
$$k = 2N\mu \times 1/2N$$

the elegant simplicity of **neutral theory**: $k = \mu$

N = small



N = large



mutation event

Why is the equilibrium substitution rate $k = \mu$?

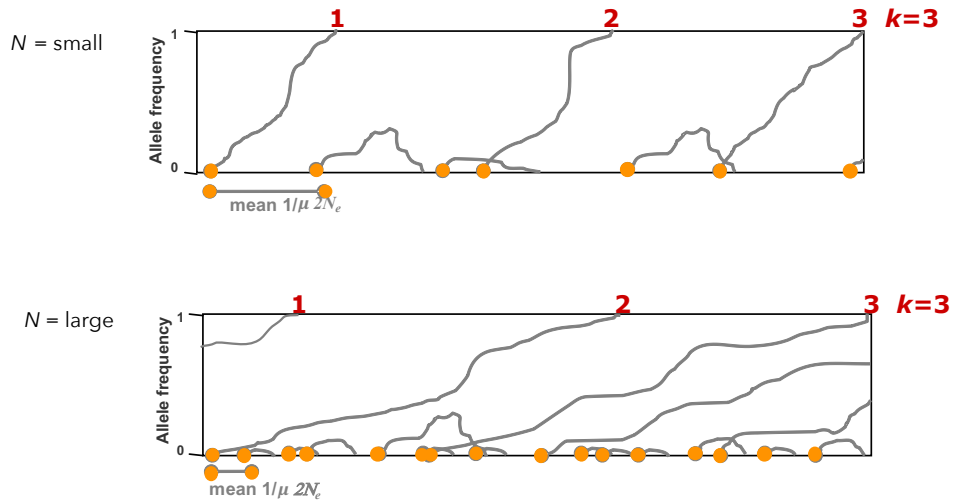
neutral theory: $k = \mu$

In words …

Small populations: lower number of new mutants each generation, but each has a higher probability of fixation

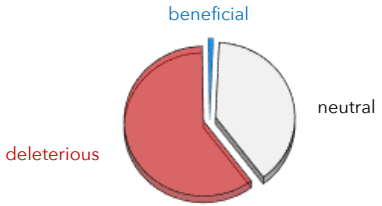
Large populations: higher number of new mutants each generation but probability of fixation is lower

Neutral Theory: precise expectations when mutation & drift are at equilibrium



neutral theory: $k = \mu$

1. evolutionary rate is independent of population size
2. constant neutral rate : "**molecular clock**"
3. evolutionary rate is inverse of functional constraint



Distribution of fitness effects (**DFE**) of **MUTATIONS** according to Kimura's Neutral Theory

lethal & strongly deleterious mutations:

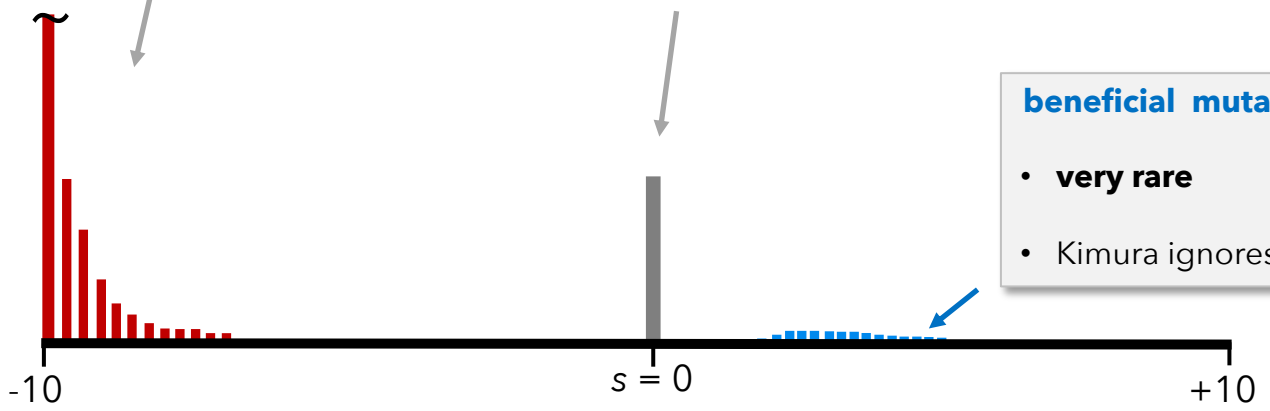
- rapidly removed by natural selection
- **never observed** in natural populations
- Kimura ignores them

neutral mutations = vast majority of:

- polymorphism
- species divergence

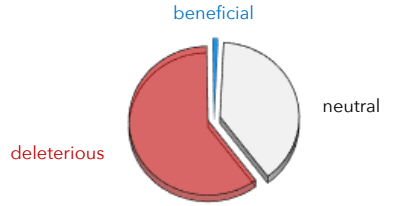
beneficial mutations:

- **very rare**
- Kimura ignores them



selection coefficients (s) for new mutations

neutral theory: $k = \mu$



Distribution of fitness effects (**DFE**) of **MUTATIONS** according to Kimura's Neutral Theory

lethal & strongly deleterious mutations:

- rapidly removed by natural selection
- **never observed** in natural populations
- Kimura ignores them

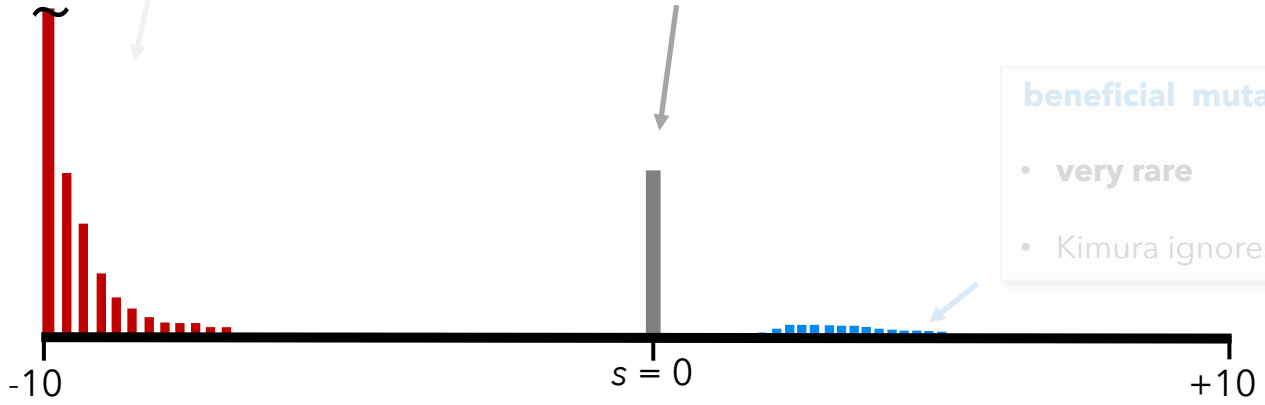
neutral theory: $k = \mu$

neutral mutations = vast majority of:

- polymorphism
- species divergence

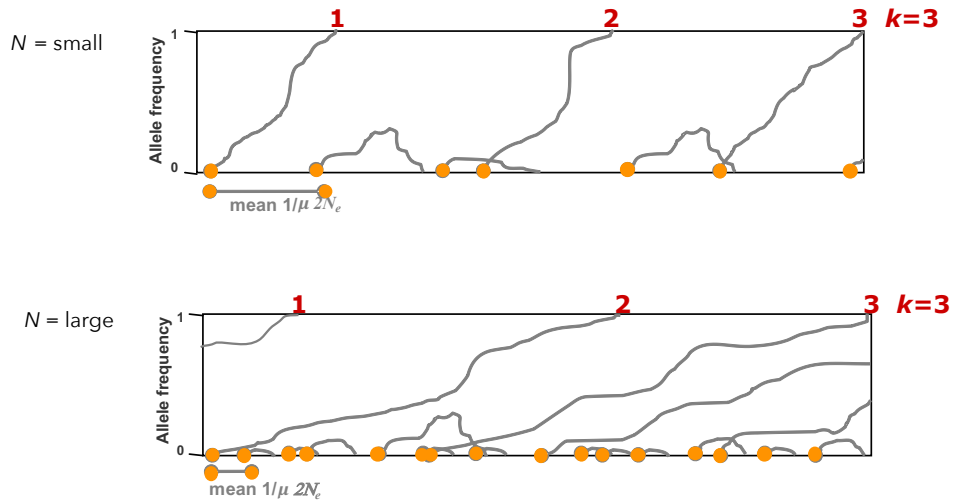
beneficial mutations:

- **very rare**
- Kimura ignores them



selection coefficients (s) for new mutations

Neutral Theory: precise expectations when mutation & drift are at equilibrium



neutral theory: $k = \mu$

1. evolutionary rate is independent of population size
2. constant neutral rate : "**molecular clock**"
3. evolutionary rate is inverse of functional constraint

Distribution of fitness effects (**DFE**) according to Kimura's Neutral Theory

neutral theory: $k = \mu$

The **ratio** of these determines the rate of evolution

lethal & strongly deleterious mutations:

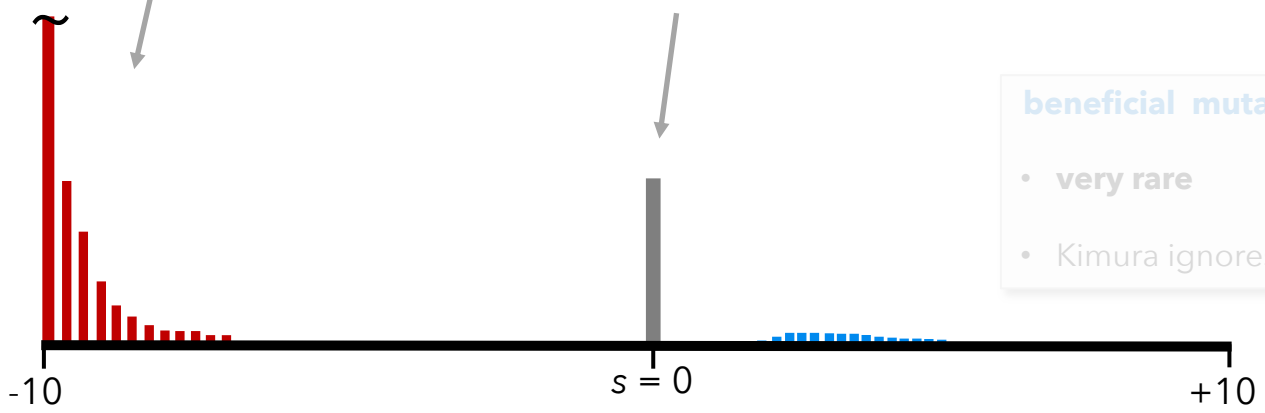
- rapidly removed by natural selection
- **never observed** in natural populations
- Kimura ignores them

neutral mutations = vast majority of:

- polymorphism
- species divergence

beneficial mutations:

- **very rare**
- Kimura ignores them



selection coefficients (s) for new mutations

Distribution of fitness effects (**DFE**) according to Kimura's Neutral Theory

neutral theory: $k = \mu$

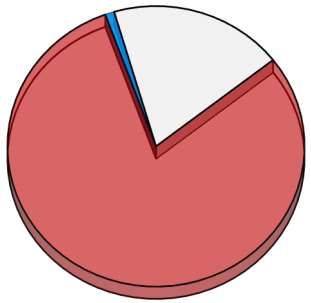
The **ratio** of these determines the rate of evolution

lethal & strongly deleterious mutations:

- rapidly removed by natural selection
- **never observed** in natural populations
- Kimura ignores them

neutral mutations = vast majority of:

- polymorphism
- species divergence



Lower-rate gene:

- more sites functional
- more mutations removed by selection

Distribution of fitness effects (**DFE**) according to Kimura's Neutral Theory

neutral theory: $k = \mu$

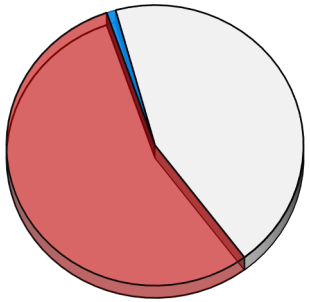
lethal & strongly deleterious mutations:

- rapidly removed by natural selection
- **never observed** in natural populations
- Kimura ignores them

The **ratio** of these determines the rate of evolution

neutral mutations = vast majority of:

- polymorphism
- species divergence



High-rate gene:

- more neutral sites
- more mutations fixed by drift

neutral theory predicts:

*The evolutionary rate is **inverse** of functional constraint.*

Conceptual models that explain genetic variation:

1. Neo-Darwinism

natural selection dominates

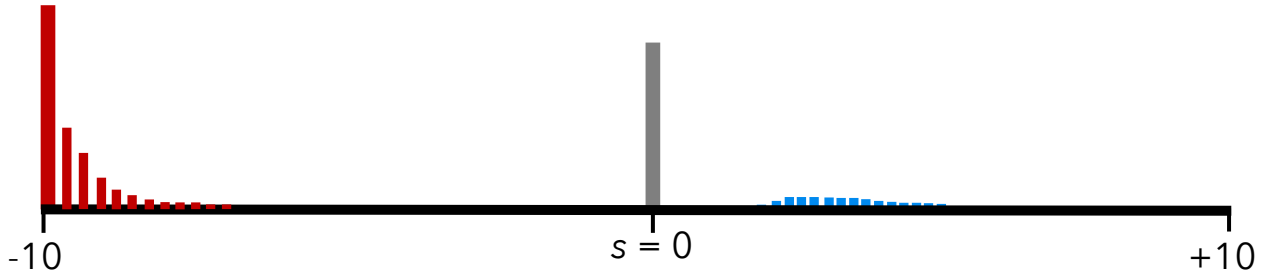
2. Neutral Theory

genetic drift dominates

3. Nearly Neutral

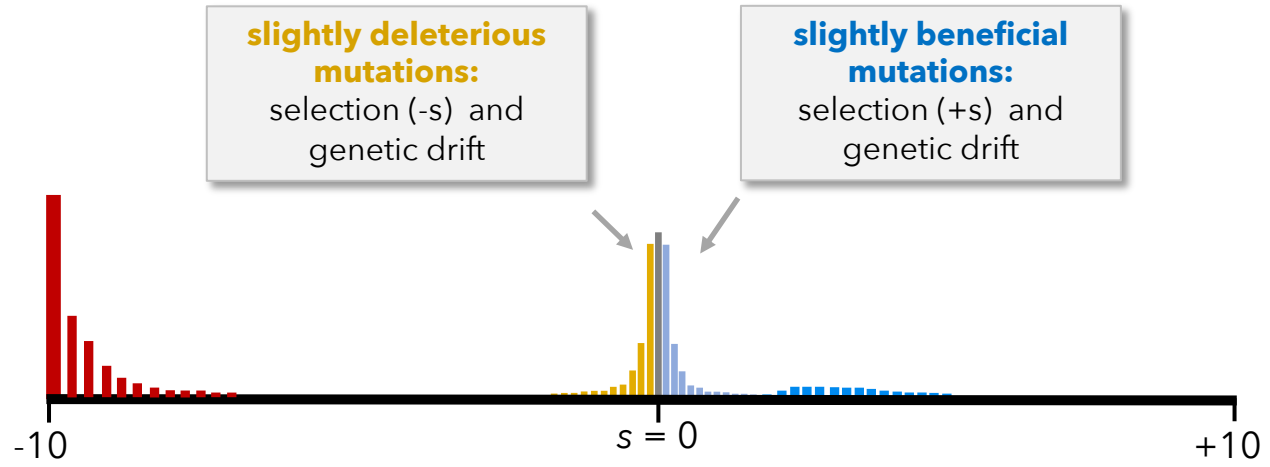
drift and selection **interact**

DFE for Kimura's Neutral Theory



evolution is modelled as an "*all or nothing affair*"

DFE for Ohta's Nearly-Neutral Theory



evolution is modelled as an **interaction** between genetic drift and natural selection

Distribution of fitness effects (DFE): broad importance to evolutionary biology

Evolutionary Importance:

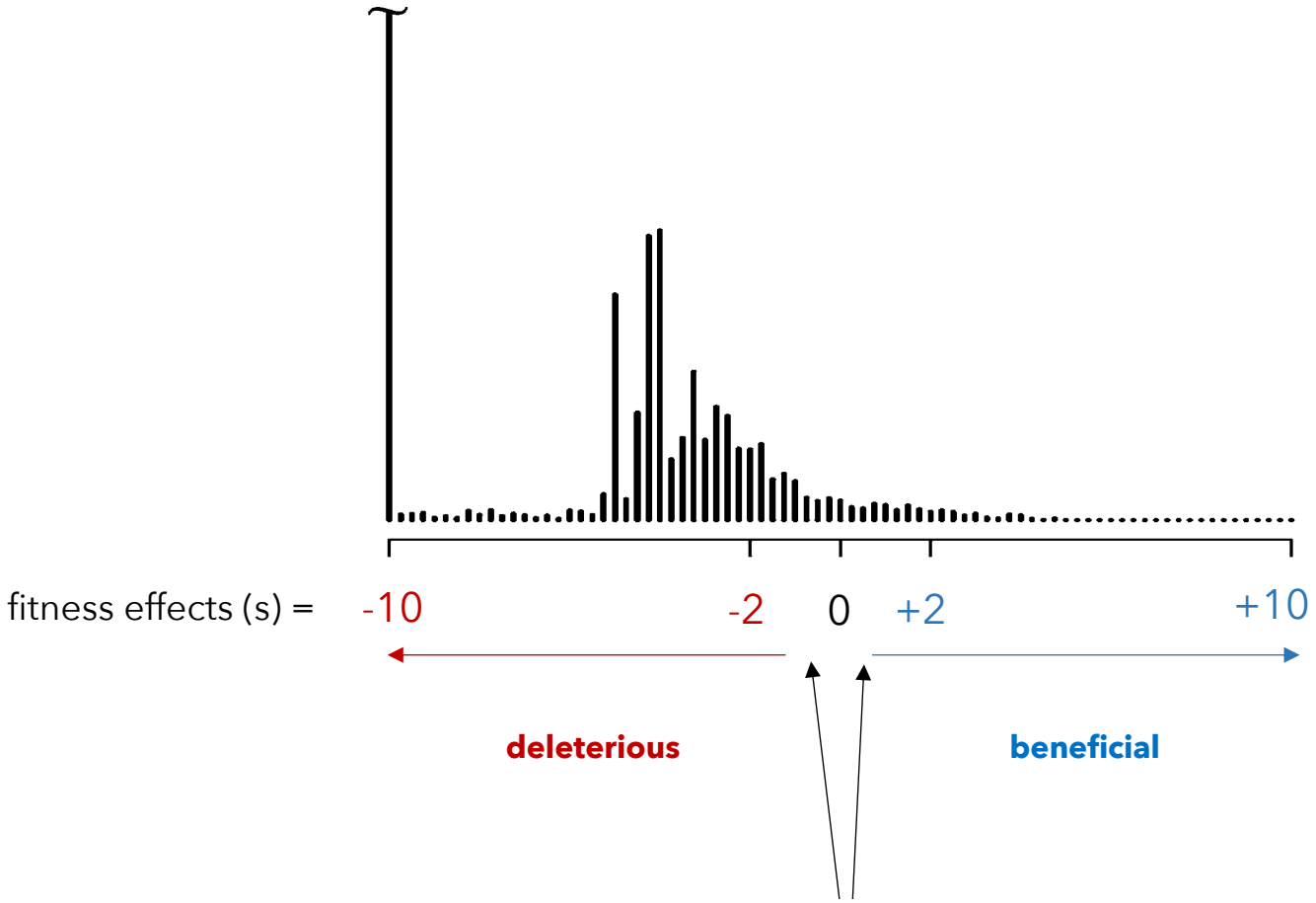
- mutations are ultimate source of variation
- rate of evolution
- species adaptation
- mutational load (genome decay & reduced survival)

Inference DFEs:

- longstanding goal
- hard to estimate
- much variation in observed DFEs

Distribution of fitness effects: non-synonymous mutations in viral PB2 gene

example: Tamuri et al. (2012) Genetics. 190:1101-1115.



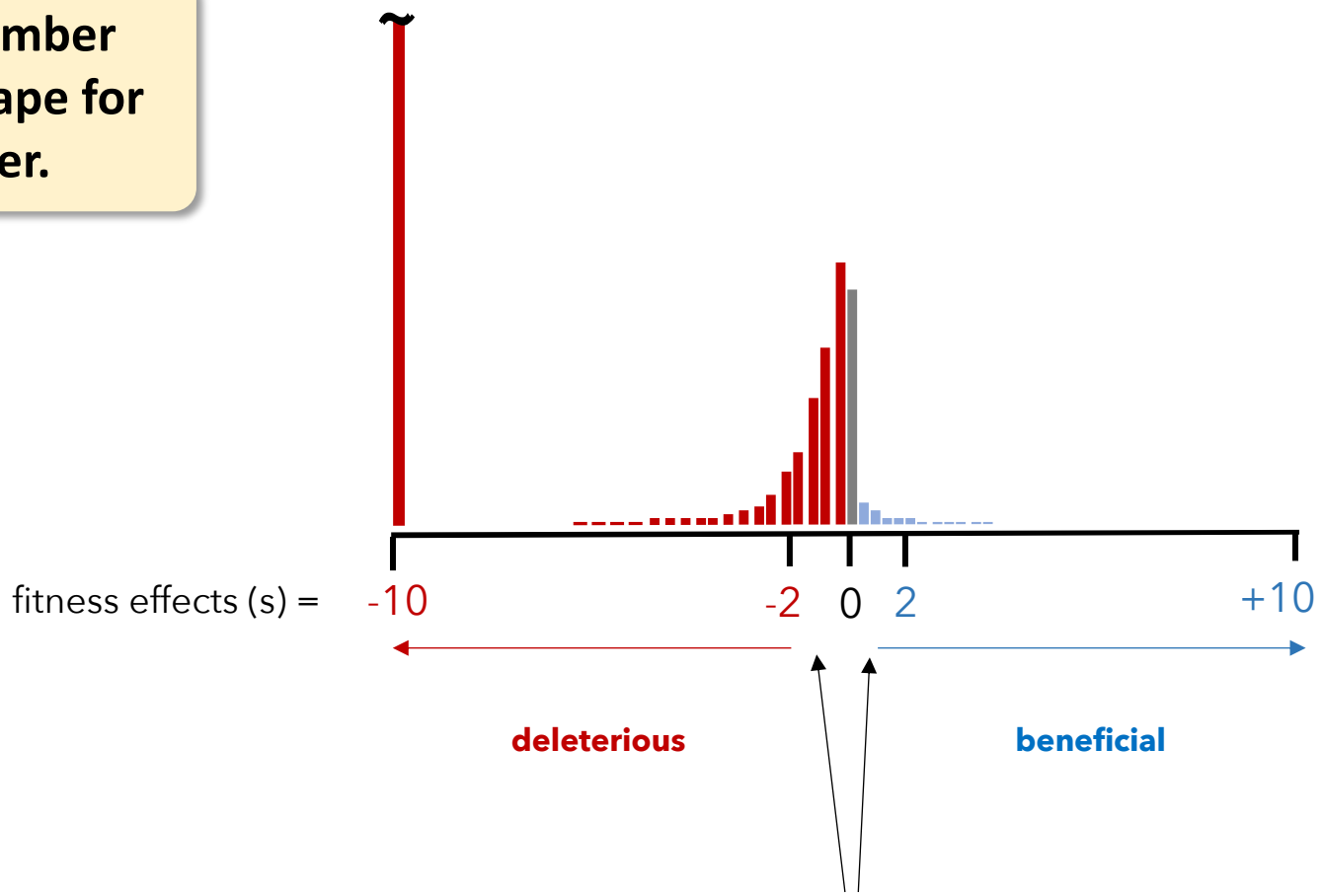
nearly neutral mutations:
drift & selection **interact**

features:

- bimodal
- deleterious: many
- beneficial: fewer
- **nearly neutral: spectrum**

Distribution of fitness effects: generalized compilation of inferences

Remember this shape for later.

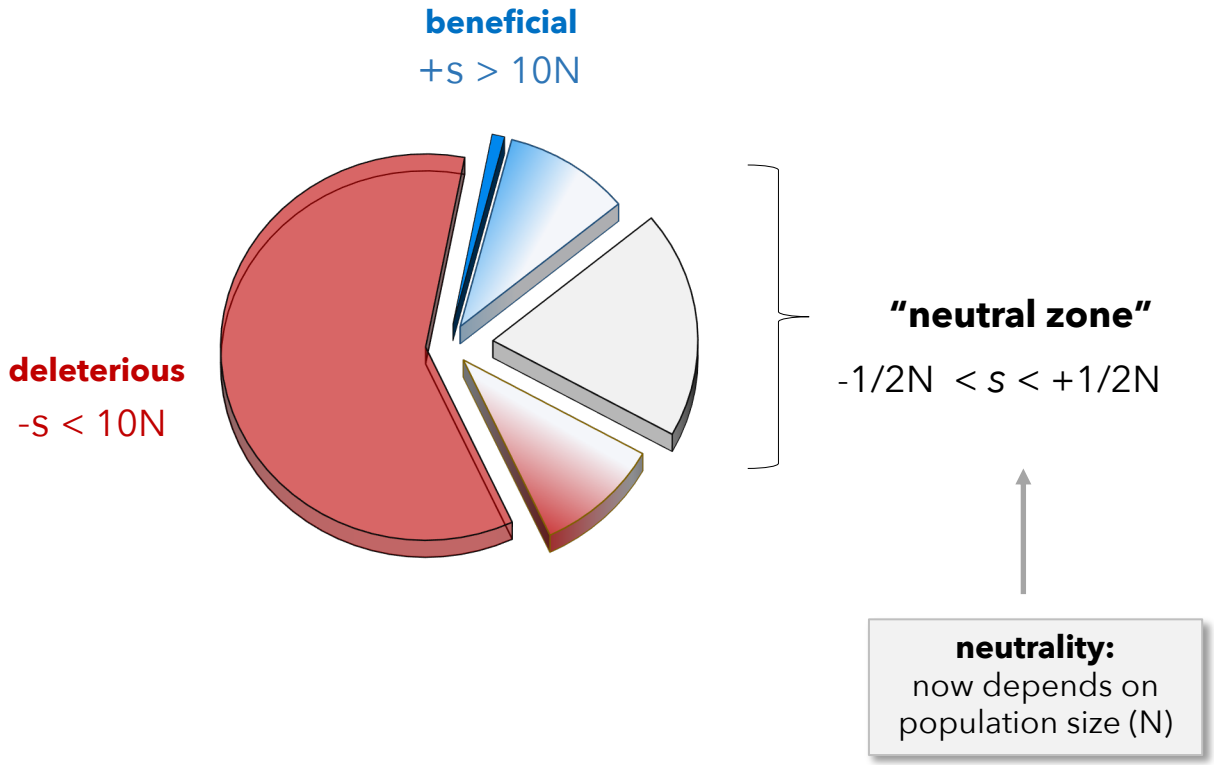


nearly neutral mutations:
drift & selection **interact**

features:

- bimodal
- deleterious: many
- beneficial: fewer with exponential character
- **nearly neutral: spectrum**

3. Nearly Neutral



drift and selection **interact**

3. Nearly Neutral

drift and selection **interact**

OHTA EXTENDS THE SELECTION AND NEUTRAL MODELS

Ohta adds evolution of "*nearly neutral*" mutations:

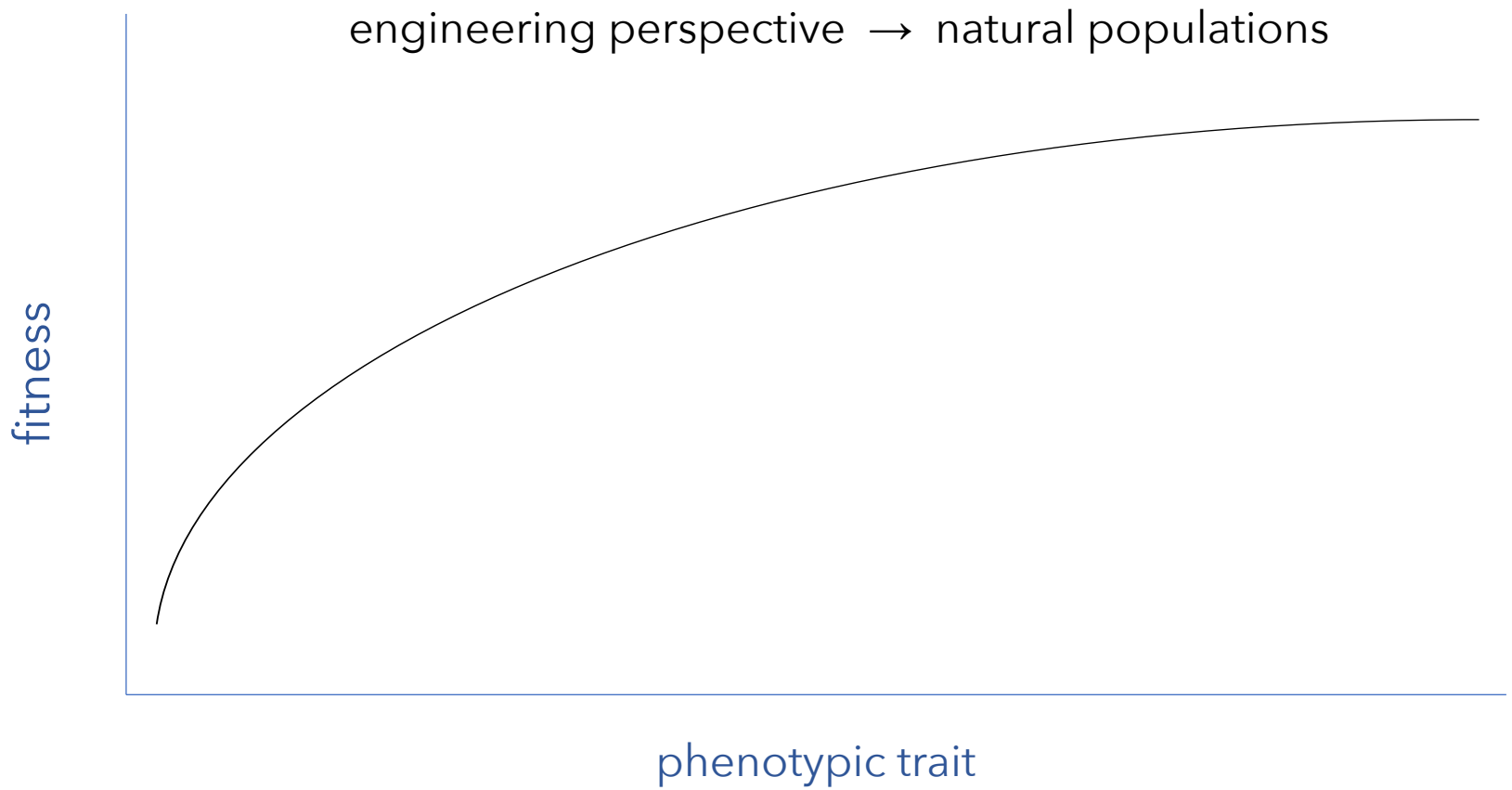
- small populations:
  - larger "neutral space"
  - more mutations are "effectively neutral"
  - more mutations evolve by genetic drift
- large populations:
  - smaller "neutral space"
  - deleterious: mostly eliminated by natural selection
  - beneficial: fixed more frequently than by drift (but fixation is not certain)

now… molecular clock unlikely (the rate of evolution is affected by changes in N)

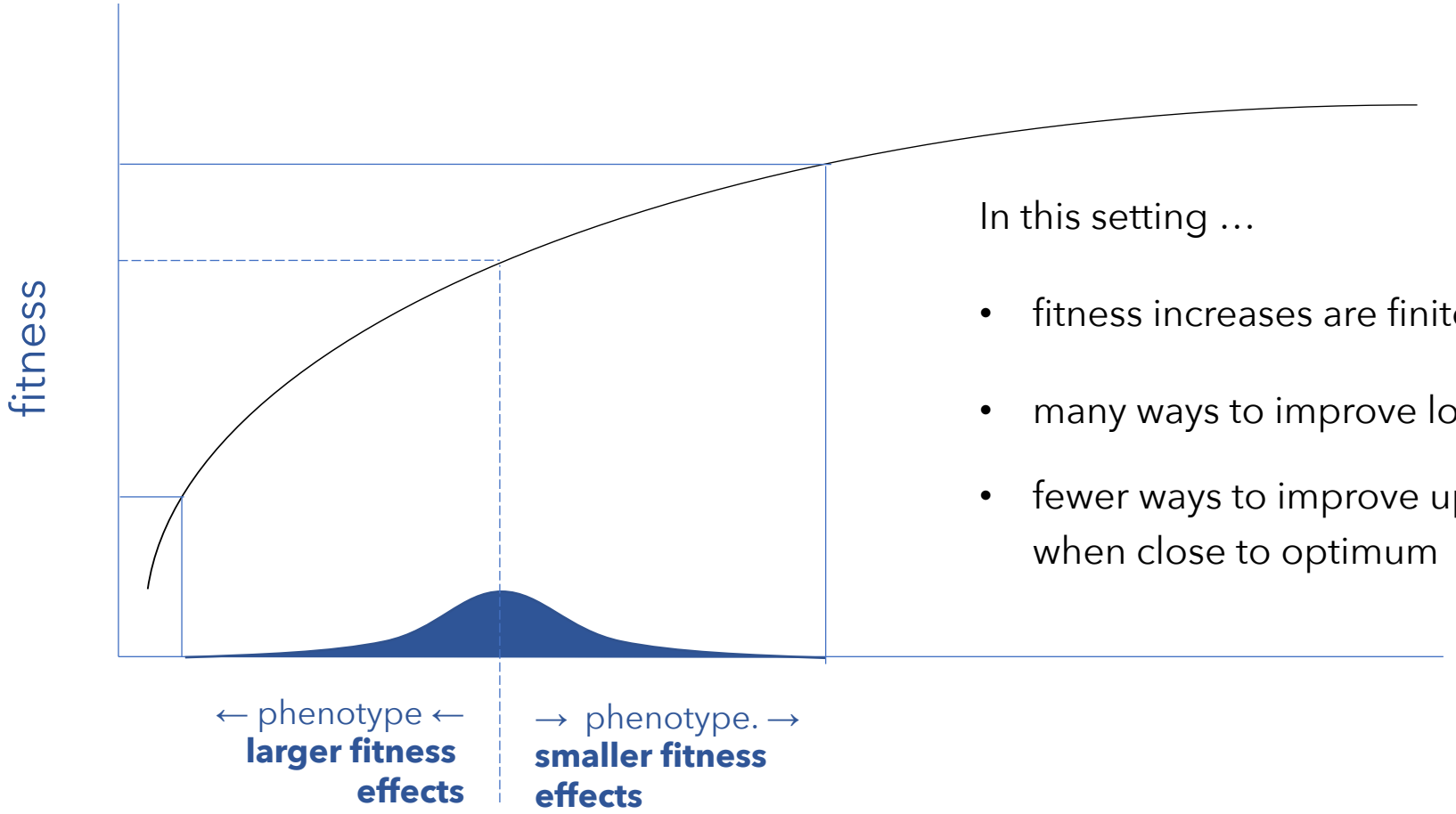
Selective implications of near neutrality

1. rate slows as population becomes adapted
2. population approaches an equilibrium
3. population reaches a state of “detailed balance”

concave (saturating) fitness curve



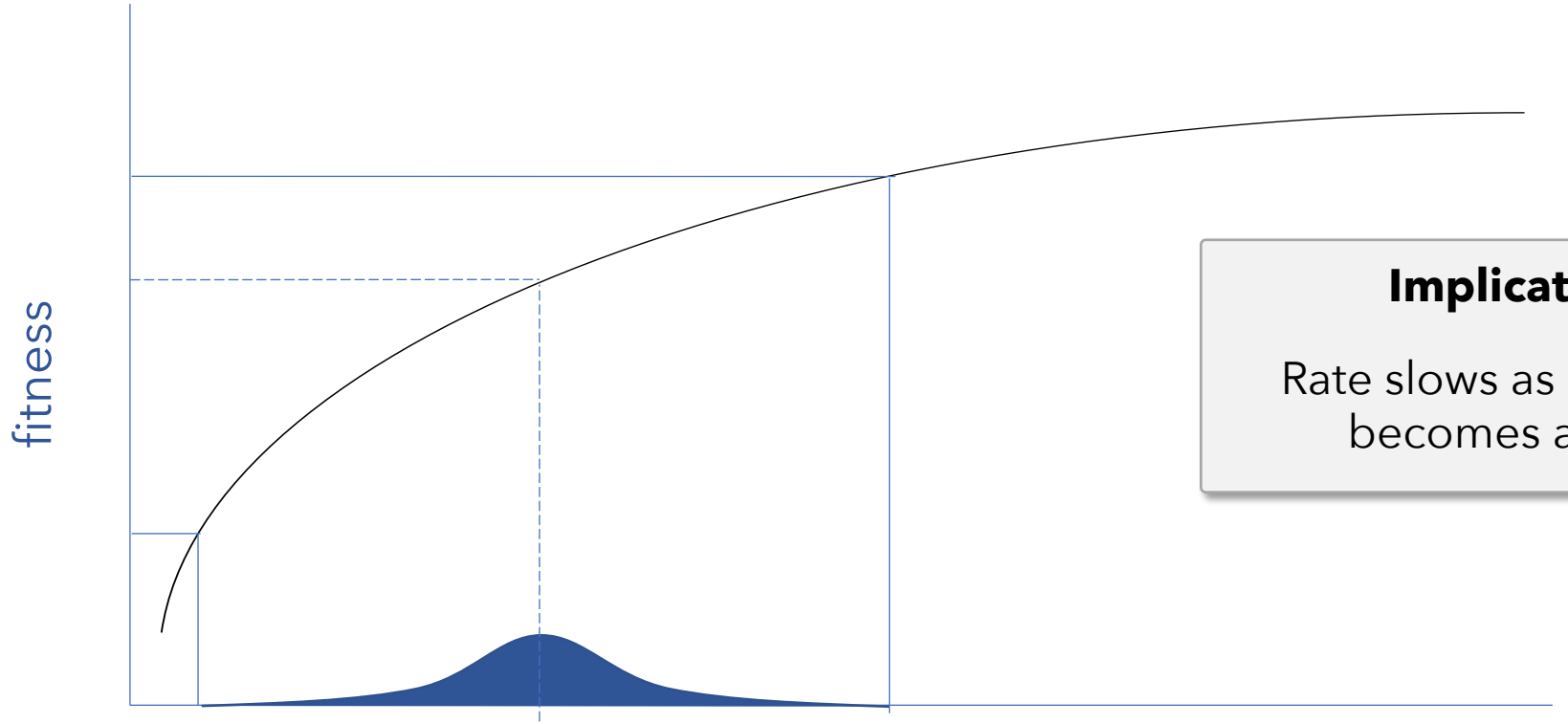
concave (saturating) fitness curve



In this setting …

- fitness increases are finite
- many ways to improve low fitness
- fewer ways to improve upon higher fitness when close to optimum

Implications of nearly neutral theory



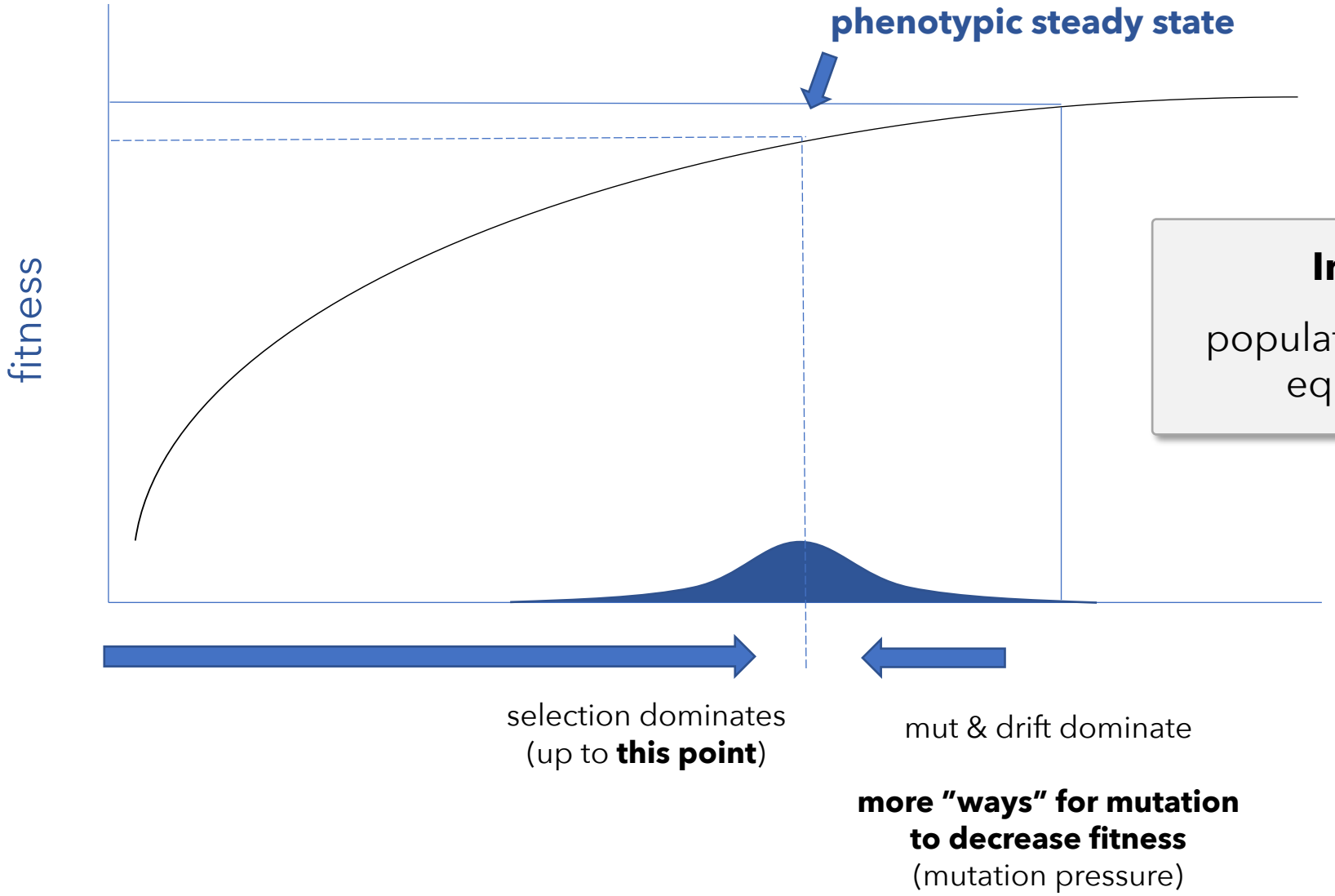
Implication 1:

Rate slows as population becomes adapted

- selection dominates

- selection less effective
- selection **interacts** with drift
- population size matters (drift stronger in smaller populations)

Implications of nearly neutral theory

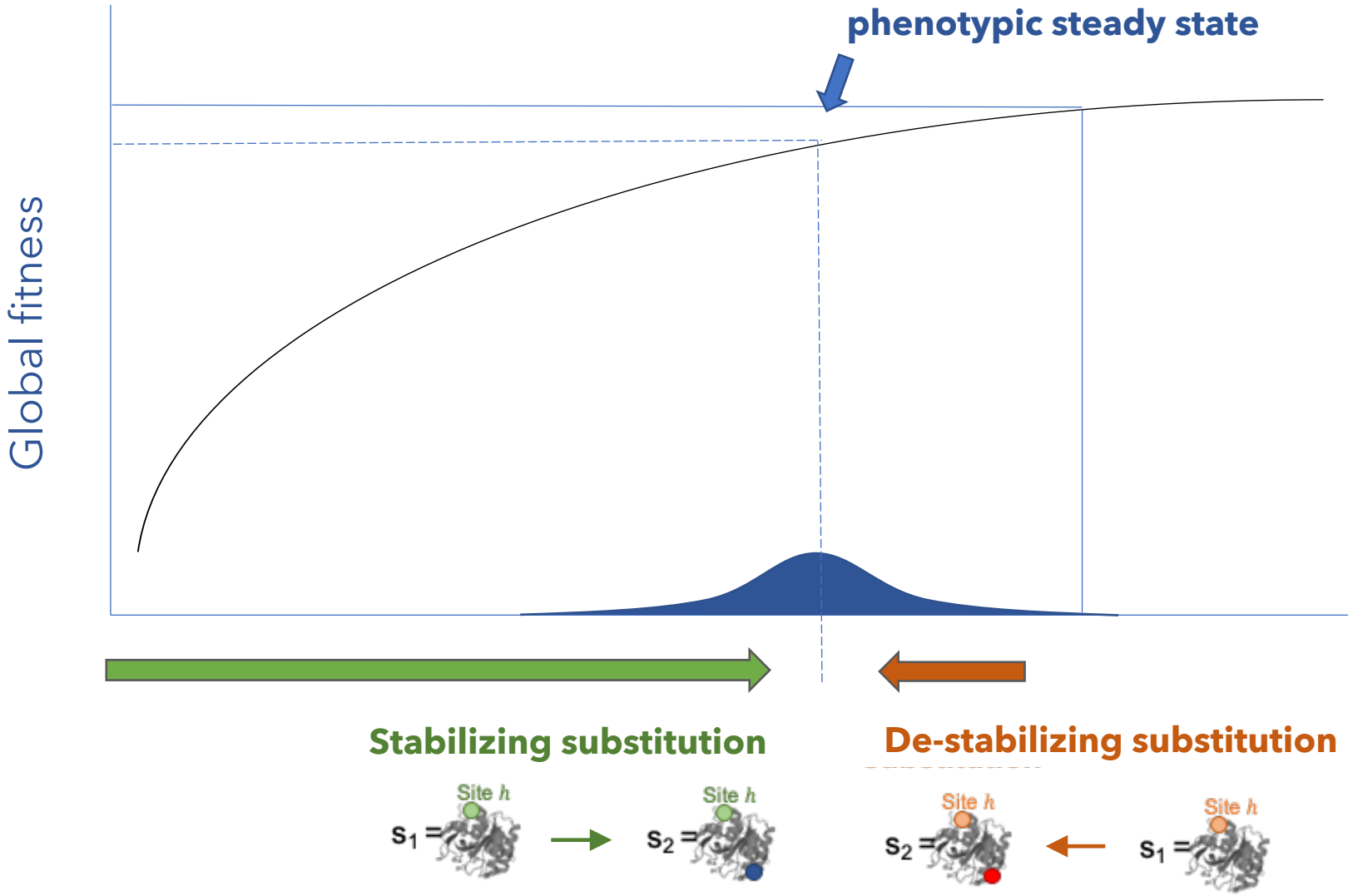


Implication 2:

population approaches an equilibrium point

An now, an actual molecular evolutionary process…

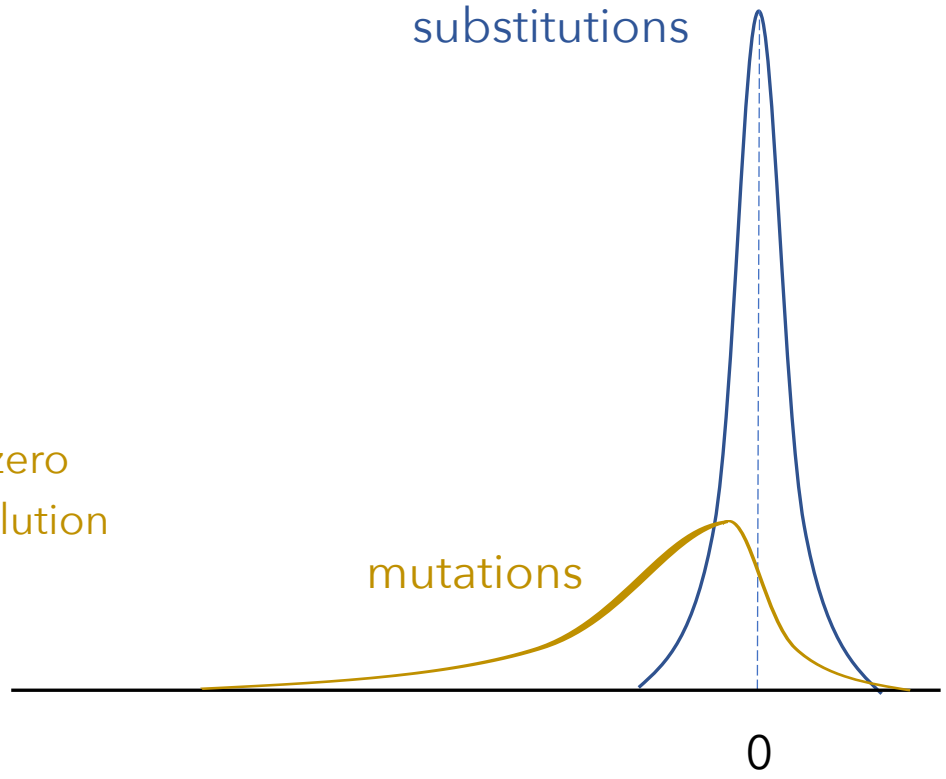
example: **epistasis and protein stability**



Implications of nearly neutral theory

DFEs:

- asymmetric around zero
- can change with evolution



equilibrium point: average affect on phenotype is **balanced** such that beneficial and deleterious substitutions of the same *absolute effect* have equal substitution rates

Implication 3:

evolutionary *process* reaches a state of "detailed balance"

This is why some forms of Nearly-Neutral models are sometimes called "**steady state models**" or "**balance mutation models**"

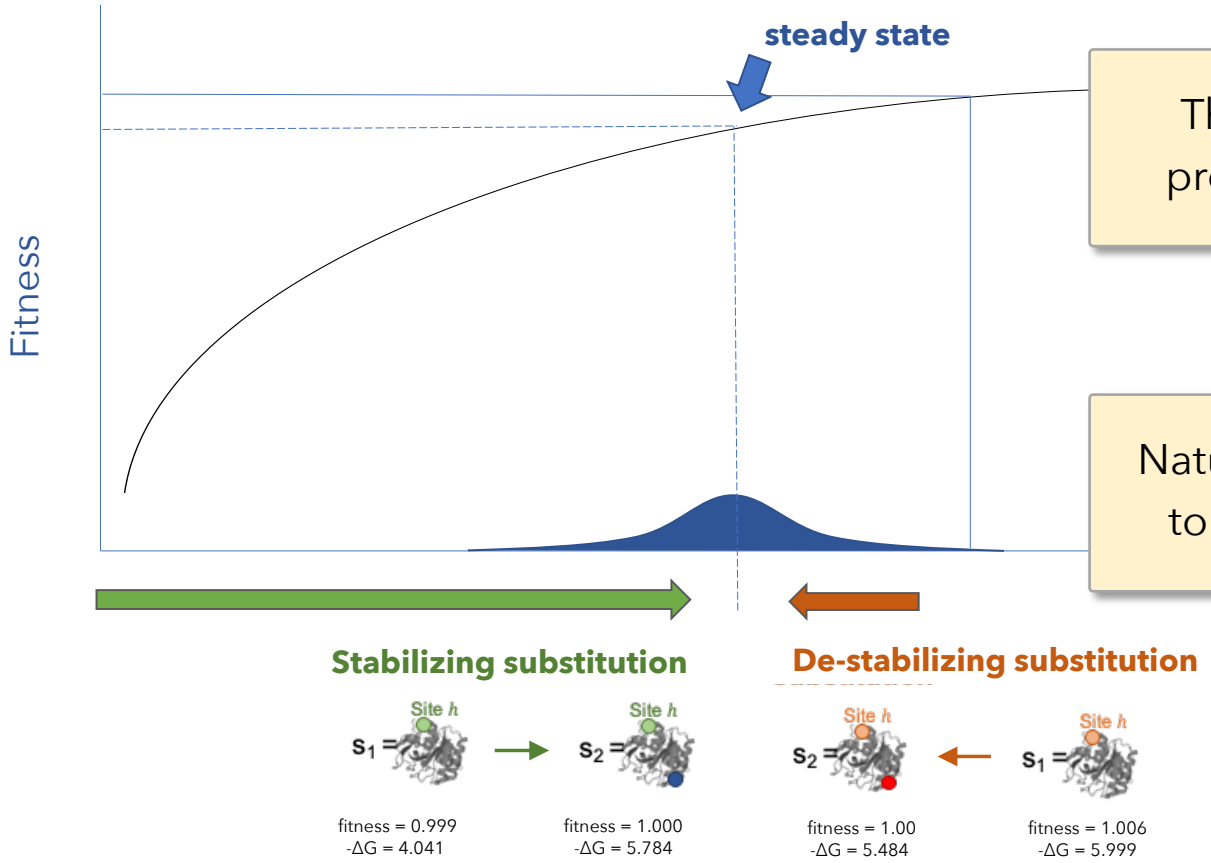
- Hartl DL, Taubes CH. Compensatory nearly neutral mutations: selection without adaptation. J Theor Biol. **1996**. 182(3):303-309.

- Sella G, Hirsh AE. The application of statistical physics to evolutionary biology. Proceedings of the National Academy of Sciences. **2005**. 102(27):9541-9546.

- Razeto-Barry P, Díaz J, Vásquez RA. The nearly neutral and selection theories of molecular evolution under the fisher geometrical framework: substitution rate, population size, and complexity. Genetics. **2012**. 191(2):523-534.

- Jones CT, Youssef N, Susko E, Bielawski JP. Shifting balance on a static mutation–selection landscape: a novel scenario of positive selection. Molecular biology and evolution. **2016**. 34(2):391-407.

The equilibrium phenotype is NOT the most fit type.
(**adaptation** ≠ optimal "engineering" state)



The "*marginal stability*" of natural proteins is **NOT an adaptive state**

Natural selection plays important role to prevent "mutational meltdown".

Further reading on **stability-mediated epistasis** and **protein evolution**…

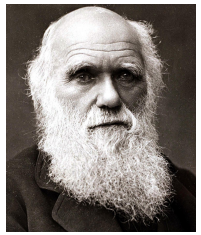
- Sella G, Hirsh AE. The application of statistical physics to evolutionary biology. Proceedings of the National Academy of Sciences. **2005**. 102(27):9541-9546.

- Goldstein RA. The evolution and evolutionary consequences of marginal thermostability in proteins. Proteins: Structure, Function, and Bioinformatics. **2011**. 79(5):1396-1407.

- Pollock DD, Thiltgen G, Goldstein RA. Amino acid coevolution induces an evolutionary Stokes shift. Proceedings of the National Academy of Sciences. **2012.** 109(21):E1352-9.

- Youssef N, Susko E, Roger AJ, Bielawski JP. Evolution of amino acid propensities under stability-mediated epistasis. Molecular Biology and Evolution. **2022**. 39(3):msac030.

Summary



1. Neo-Darwinism:
    - almost everything is adaptive (*too strong*)
    - evolution “seeks optimality” (promotes an engineering perspective)
    - remains THE evolutionary mechanism for *origin of adaptations*

2. Neutral theory:
    - elegant simplicity
    - assumes simplistic DFE
    - predictions are correct for effectively neutral mutations
    - many predictions serve as “principles of evolution”

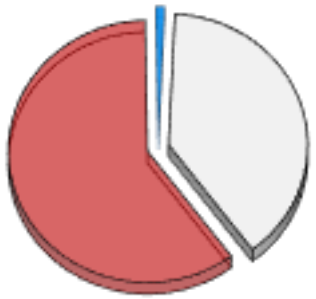
3. Nearly-neutral theory:
    - predicts more complex evolutionary dynamics
    - depends on populations size (unlike neutral theory)
    - some predictions closer to natural populations that neutral theory
    - predicts equilibrium where phenotype is not necessarily optimal
    - natural selection acts to balance “mutational load” on fitness (*maintenance*)

An index of the intensity of natural selection for proteins

Neutral and Nearly-neutral theory

Evolutionary rate depends on intensity of selection

Neutral theory:
independent of N



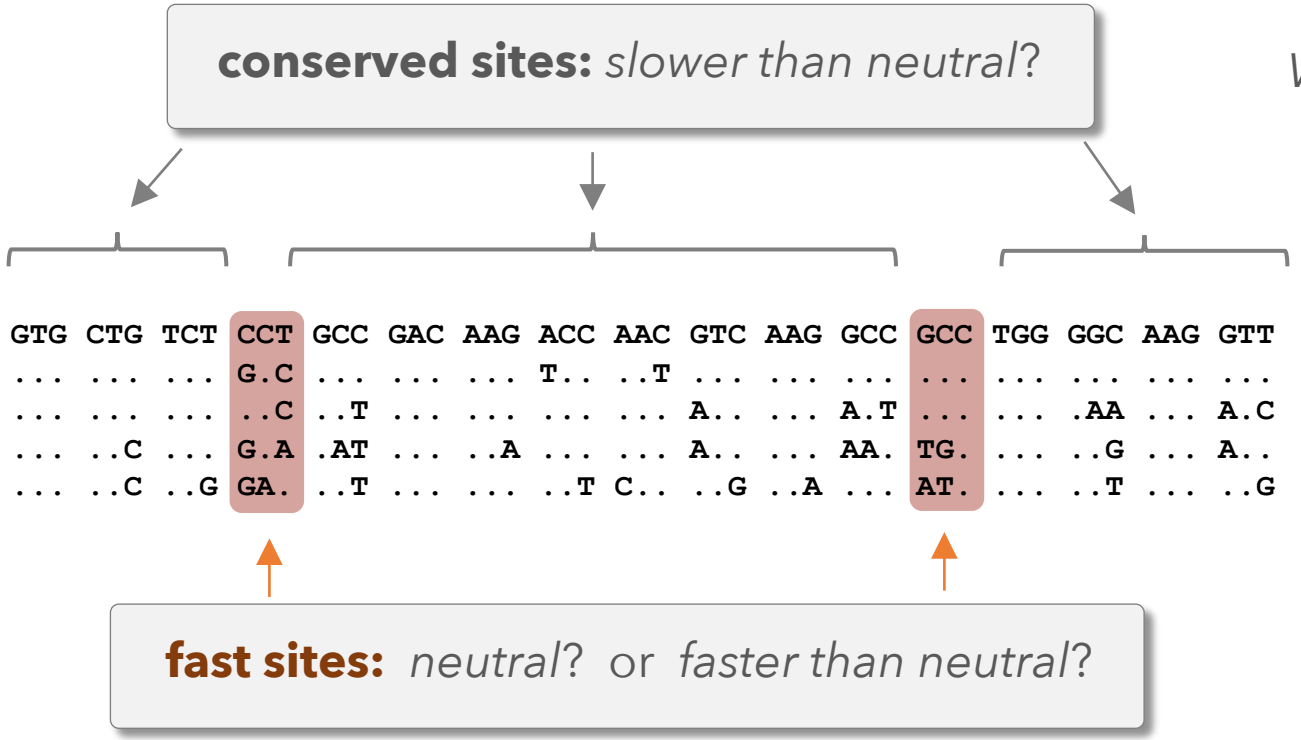
Nearly-neutral theory:
depends on N

Let’s apply these ideas to individual sites…

1.) selectively constrained:
- neutral space < 100%
- rate < strictly neutral

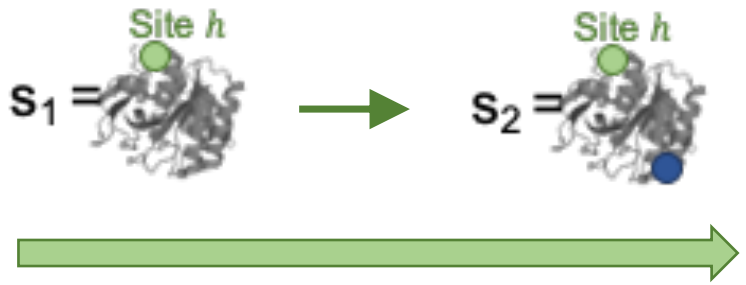
2.) strictly neutral:
- 100% neutral space
- rate = neutral rate

3.) adaptive evolution:
- large adaptive space
- rate > neutral rate (?)

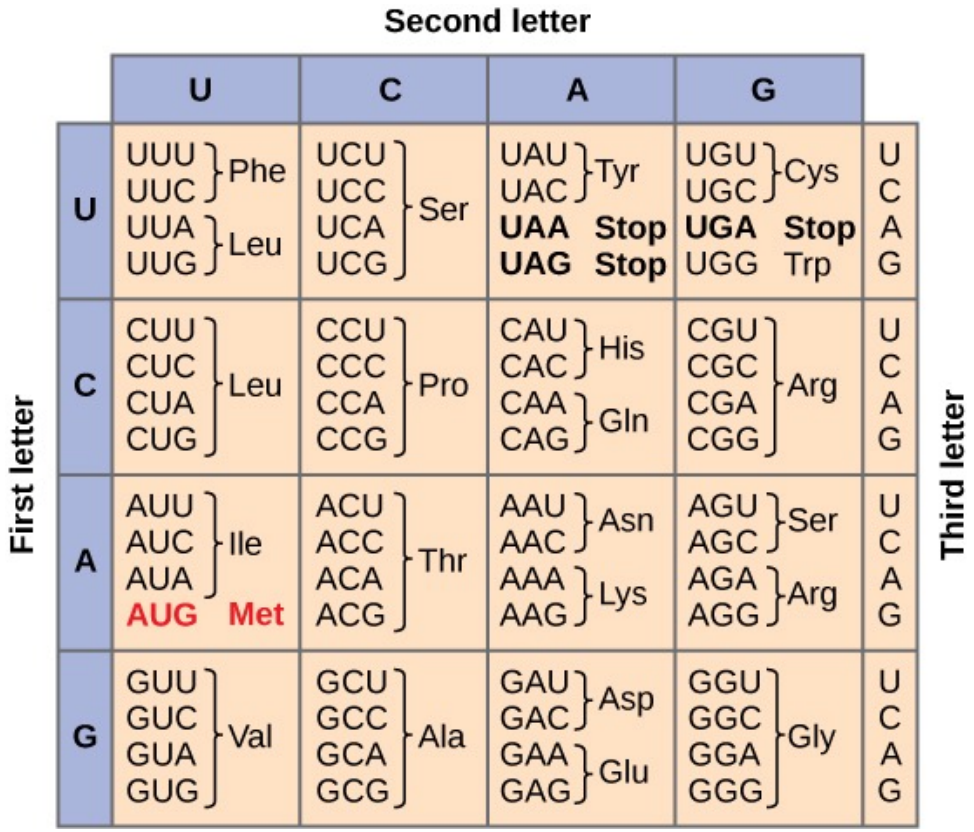


Question:
What is the neutral rate?

Proteins have a "*built in ruler*" for their own neutral rate of molecular evolution!



Genetic code



all possible mutations → **two types**

synonymous (**S**): no change to protein

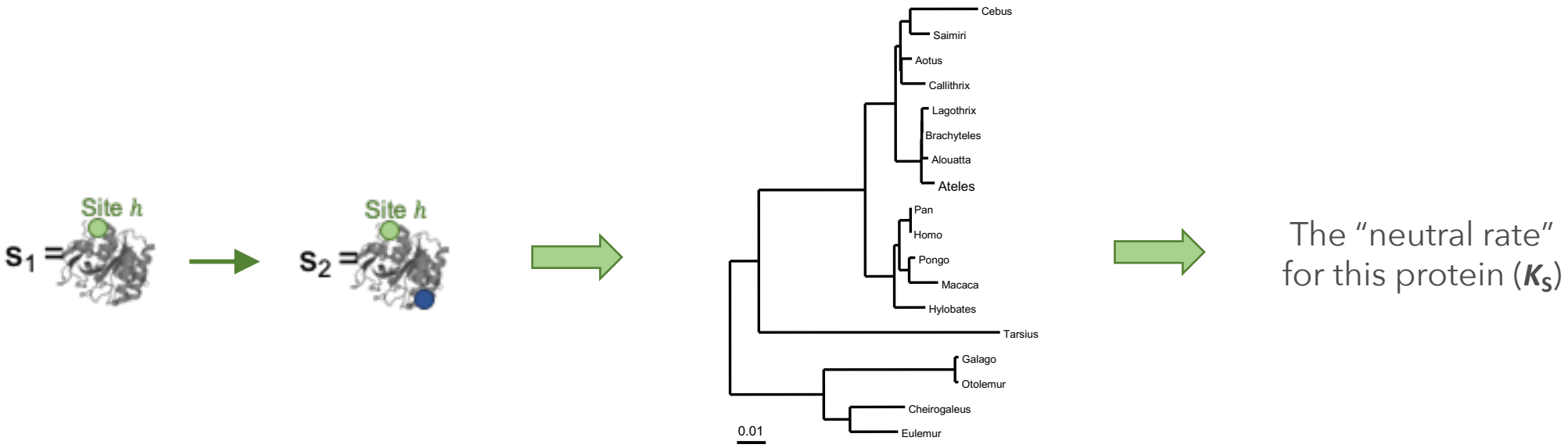
- no effect on protein
- selectively neutral
- rate = neutral rate (*w.r.t. protein evolution*)

non-synonymous (**N**): changes the amino acid composition of protein

- changes AA of protein
- **deleterious**, or neutral, or **positive**
- rate depends on intensity of selection

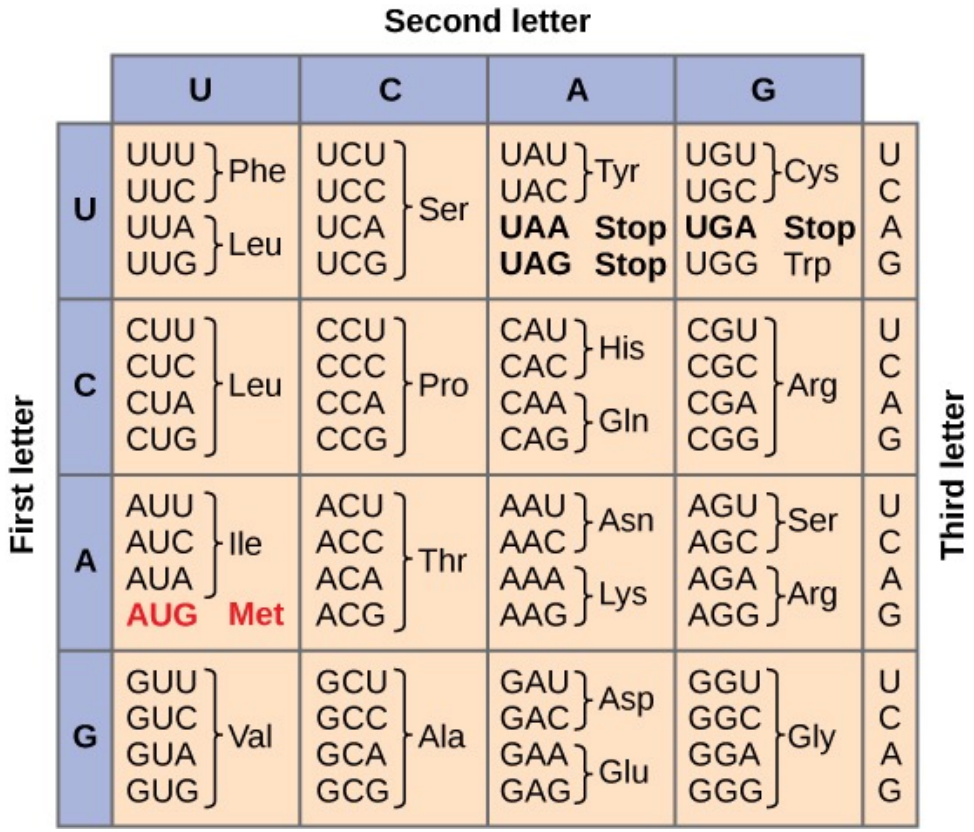
"built in ruler" = **synonymous substitution rate**

The rate at which that proteins would have evolved if it had been 100% free from selection (*at the protein level*).



The "neutral rate" for this protein (K_S)

Genetic code



all possible mutations → **two types**

synonymous (**S**): no change to protein

- no effect on protein
- selectively neutral
- rate = neutral rate
- neutral rate = K_S

non-synonymous (**N**): changes the amino acid composition of protein

- changes AA of protein
- deleterious, or neutral, or positive
- rate depends on intensity of selection
- purifying selection: $K_N < K_S$

Kimura (1983) : the **rate ratio** is an index of selection intensity

rate ratio	mode	example
$K_N/K_S < 1$	**purifying (negative) selection**	**histones**
$K_N/K_S = 1$	**neutral evolution**	**Pseudogenes**
$K_N/K_S > 1$	**diversifying selection**	**MHC, Lysin**

NOTATION:

$K_S = d_S$: number of synonymous substitutions per synonymous site

$K_N = d_N$: number of nonsynonymous substitutions per nonsynonymous site

$$\omega = \frac{dN}{dS}$$

This is the "***omega ratio***" in codon models

Note: not calling this "positive" selection (yet)

1.) selectively constrained:

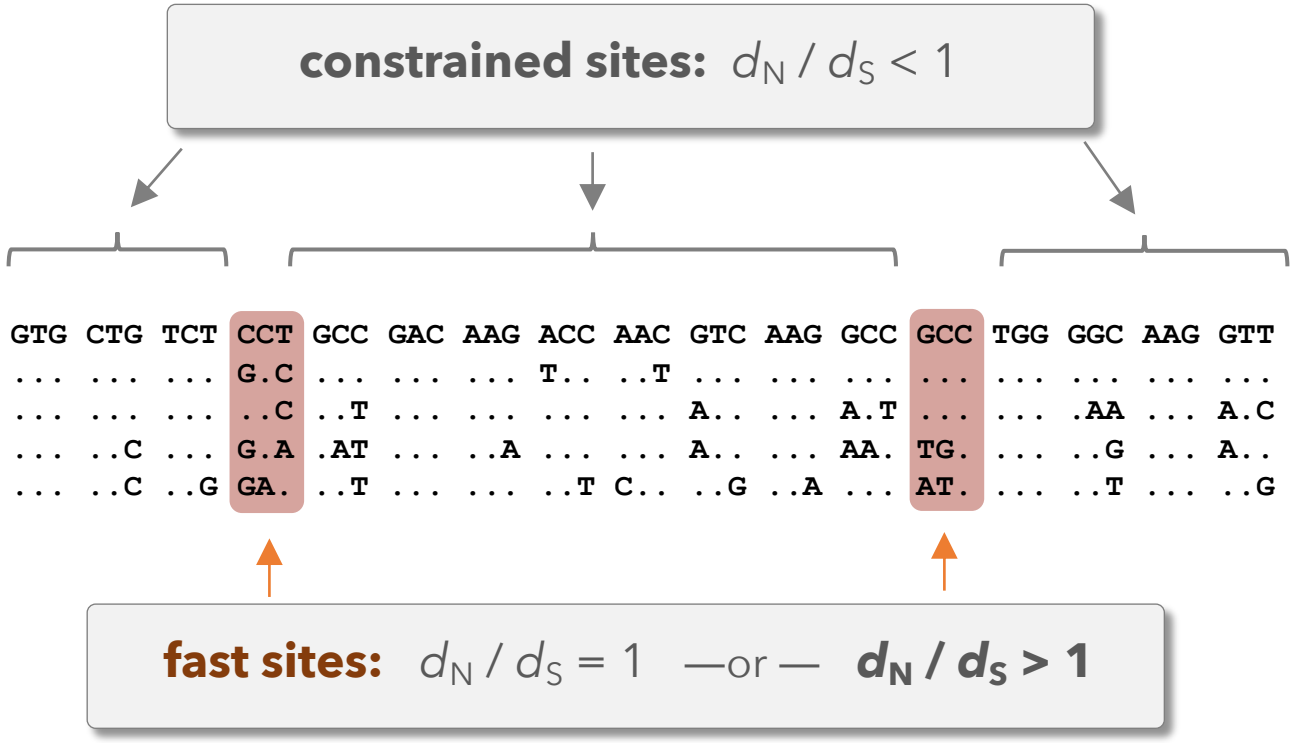
$d_N / d_S < 1$

2.) strictly neutral:

$d_N / d_S = 1$

3.) adaptive evolution:

$d_N / d_S > 1$



a useful perspective…

Consequences of Stability-Induced Epistasis for Substitution Rates

Noor Youssef,*[1,2] Edward Susko,[2,3] and Joseph P. Bielawski[1,2,3]

[1]Department of Biology, Dalhousie University, Halifax, Nova Scotia, Canada
[2]Centre for Genomics and Evolutionary Bioinformatics, Dalhousie University, Halifax, Nova Scotia, Canada
[3]Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia, Canada

***Corresponding author**: E-mail: n.youssef@dal.ca.
Associate editor: Jeffrey Thorne

Abstract

Do interactions between residues in a protein (i.e., epistasis) significantly alter evolutionary dynamics? If so, what consequences might they have on inference from traditional codon substitution models which assume site-independence for the sake of computational tractability? To investigate the effects of epistasis on substitution rates, we employed a mechanistic mutation-selection model in conjunction with a fitness framework derived from protein stability. We refer to this as the stability-informed site-dependent (S-SD) model and developed a new stability-informed site-independent (S-SI) model that captures the average effect of stability constraints on individual sites of a protein. Comparison of S-SI and S-SD offers a novel and direct method for investigating the consequences of stability-induced epistasis on protein evolution. We developed S-SI and S-SD models for three natural proteins and showed that they generate sequences consistent with real alignments. Our analyses revealed that epistasis tends to increase substitution rates compared with the rates under site-independent evolution. We then assessed the epistatic sensitivity of individual site and discovered a counterintuitive effect: Highly connected sites were less influenced by epistasis relative to exposed sites. Lastly, we show that, despite the unrealistic assumptions, traditional models perform comparably well in the presence and absence of epistasis and provide reasonable summaries of average selection intensities. We conclude that epistatic models are critical to understanding protein evolutionary dynamics, but epistasis might not be required for reasonable inference of selection pressure when averaging over time and sites.

Key words: **epistasis, dN/dS, protein stability, substitution rates, mutation-selection model, protein evolution.**

Introduction

Most proteins must fold into a native structure in which they are moderately stable before they are able to perform their biological function. Protein stability depends on the sequence of amino acids and their interactions in the folded three-dimensional structures. Because of these interactions, evolutionary selective constraints to maintain adequate stability result in epistatic dependencies between residues. Specifically, epistasis manifests as a dependency in the fitness effect of a mutation on the background protein sequence in which it arose. For example, let $f_a^h(S)$ be the fitness of the protein provided amino acid a is occupying site h in the context of background sequence S. Then, $F^h(S) = \langle f_1^h(S), \ldots, f_{20}^h(S) \rangle$ is the site-specific vector of amino acid fitness values specifying the fitness landscape at site h. Following a substitution at another position in the protein, so that the background sequence changes from S to X, the fitness of the same amino acid will subsequently change, $f_a^h(S) \neq f_a^h(X)$. Therefore, in the presence of epistatic dependencies, the fitness landscape at a site is subject to fluctuations as substitutions occur at other sites (fig. 1A). Stability constraints typically result in global epistasis, meaning that a change in the incumbent amino acid at one site induces shifts in the fitness landscapes at many, often all, other sites in the protein (Starr and Thornton 2016). Although such interdependencies inevitably occur, the magnitude and frequency of these shifts, and their impact on protein evolution, remain controversial.

Using extensive computational experiments, Pollock et al. (2012) found that stability-induced epistasis results in frequent and substantial shifts in amino acid fitness landscapes. To the contrary, Ashenberg et al. (2013) used computational and experimental approaches and reported that although stability-induced fluctuations in site-specific amino acid fitness landscapes do occur, they are relatively minor in magnitude and are therefore inconsequential with regards to long-term evolutionary dynamics. This controversy has spurred multiple follow-up experiments, finding support for both claims and little consensus (Risso et al. 2015; Shah et al. 2015; Starr et al. 2018; Ferrada 2019). It remains unclear if and how stability-induced epistasis influences protein evolution.

Models used to infer evolutionary parameters from natural protein alignments commonly assume site-independence and other simplifying assumptions (e.g., time-stationary substitution rates, and low levels of among-site rate

Article

"…***despite the unrealistic assumptions***, *traditional models perform comparably well in the presence and absence of epistasis and provide* ***reasonable summaries of average selection intensities***."

This does **NOT** mean that you can just plow ahead and *ignore the assumptions* of your models and your tests!!!