Molecular Evolution

COMP 571 - Fall 2010

Luay Nakhleh, Rice University

Outline

(1) The neutral theory

(2) Measures of divergence and polymorphism

(3) DNA sequence divergence and the molecular clock

(4) Testing the molecular clock hypothesis

(5) Testing the neutral theory null model of DNA sequence evolution

(6) Molecular evolution of loci that are not independent

(1) The Neutral Theory

* Molecular evolution involves the study of molecular sequences (DNA, RNA, protein) with the goal of elucidating the processes that cause both change and constancy among sequences over time

* A salient feature of all hypothesis tests in studies of molecular evolution is the use of null and alternative hypotheses for the patterns and rates of sequence change

(1) The Neutral Theory

* The **neutral theory** forms the basis of the most widely employed null model in molecular evolution
* The neutral theory adopts the perspective that most mutations have little or no fitness advantage or disadvantage and are therefore **selectively neutral**
* Genetic drift is therefore the primary evolutionary process that dictates the fate of newly occurring mutations

(1) The Neutral Theory

* The neutral theory null model makes two major predictions under the assumption that genetic drift alone determines the fate of new mutations:
  * The amount of **polymorphism** for sequences sampled within a population of one species
  * The degree and rate of **divergence** among sequences sampled from separate species

(1) The Neutral Theory

Polymorphism

* Polymorphism in the neutral theory is determined by the balance between genetic drift and mutation

* The frequency of each allele is a random walk between fixation and loss

* The population has genetic polymorphism at any point in time where fixation/loss haven't been reached (multiple alleles are still **segregating**)

(1) The Neutral Theory

Polymorphism

* Recall:
  * The initial frequency of an allele is also its chance of eventual fixation
  * The average times to fixation, loss, and of segregation of an allele whose initial frequency is p are, respectively,

$$\overline{T}_{fix} = -4N\frac{(1-p)\ln(1-p)}{p} \qquad \overline{T}_{loss} = -4N\frac{p\ln p}{1-p}$$

$$\overline{T}_{segregate} = p\overline{T}_{fix} + (1-p)\overline{T}_{loss} = -4N[(1-p)\ln(1-p) + p\ln p]$$

(1) The Neutral Theory

Polymorphism

* The time that a new mutation segregates in the population depends on N_e
* The chance that a new mutation goes to fixation is also directly related to N_e
* These two effects of N_e cancel each other out for neutral alleles
* The neutral theory then predicts that the rate of fixation is μ and therefore the expected time between fixations is $1/\mu$ generations

(1) The Neutral Theory

Polymorphism

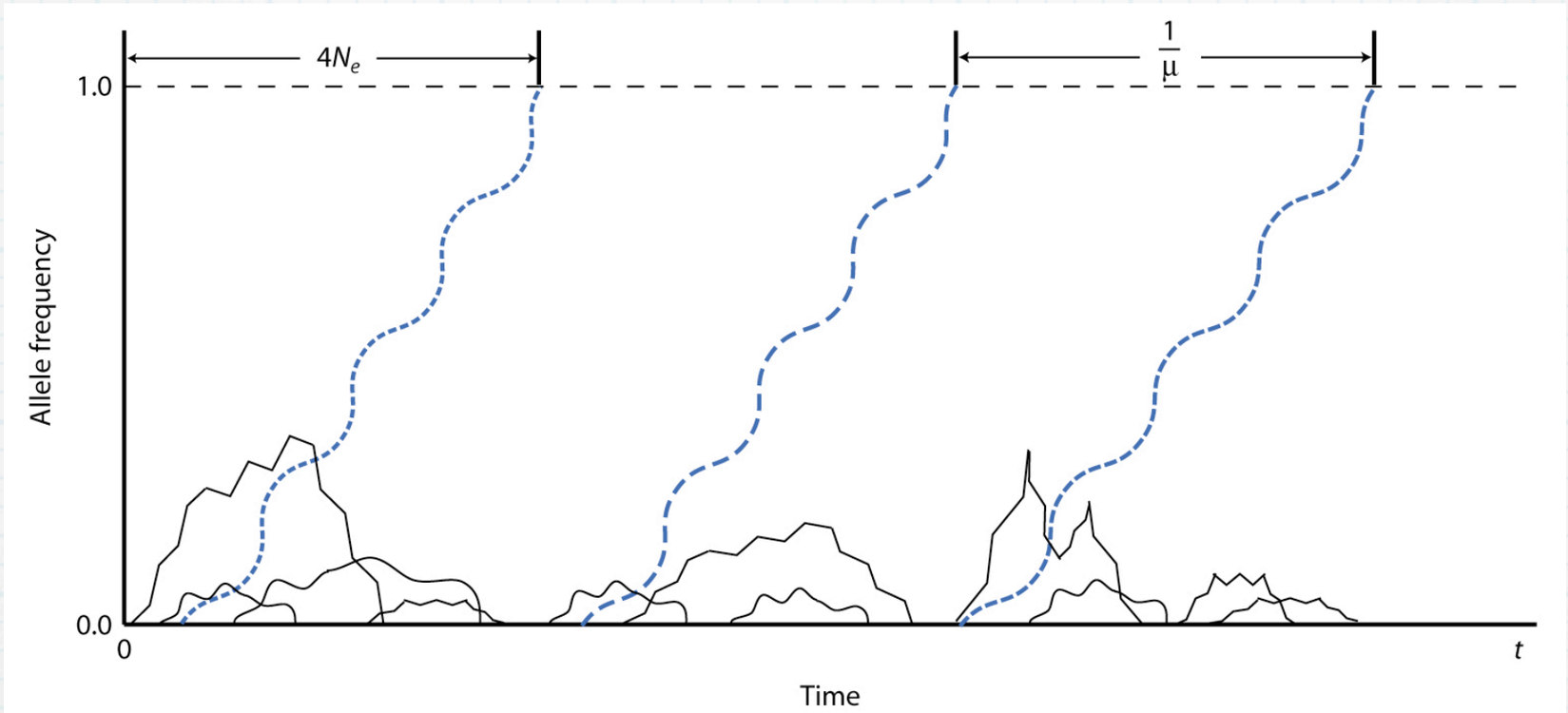


Figure 8.2 The fate of selectively neutral mutations in a population. New mutations enter the population at rate μ and an initial frequency of $\frac{1}{2N}$. Allele frequency is a random walk determined by genetic drift. The time that a new mutation segregates in the population, or the dwell time of a mutation, depends on the effective population size. However, the chance that a new mutation goes to fixation (equal to its initial frequency) is also directly related to the effective population size. These two effects of the effective population size cancel each other out for neutral alleles. The neutral theory then predicts that the rate of fixation is μ and therefore the expected time between fixations is $1/\mu$ generations. For that subset of mutations that eventually fix, the expected time from introduction to fixation is $4N_e$ generations. After Figure 3.1 in Kimura (1983a).

(1) The Neutral Theory

Polymorphism

* Another way to understand polymorphism in a population is to consider the heterozygosity in the population

* We showed that for the infinite alleles model of mutation, that combined processes of mutation and genetic drift produce equilibrium heterozygosity that depends on N_e and μ:

$$H_{equilibrium} = \frac{4N_e\mu}{4N_e\mu + 1}$$

(1) The Neutral Theory

Polymorphism

* The neutral theory prediction for polymorphism can be readily compared with polymorphism expected under positive and negative natural selection

* New mutations that are deleterious will go to loss faster than neutral mutations, while advantageous new mutations will increase in frequency to fixation

* So, a locus with new mutations that are influenced by directional natural selection should show less polymorphism than a locus with neutral mutations

(1) The Neutral Theory

Polymorphism

* Balancing selection, on the other hand, increases the time of segregation since natural selection will maintain several alleles at intermediate frequencies between fixation and loss with the result of increased levels of polymorphism in the population

(1) The Neutral Theory

directed selection →

neutral →

balancing selection →

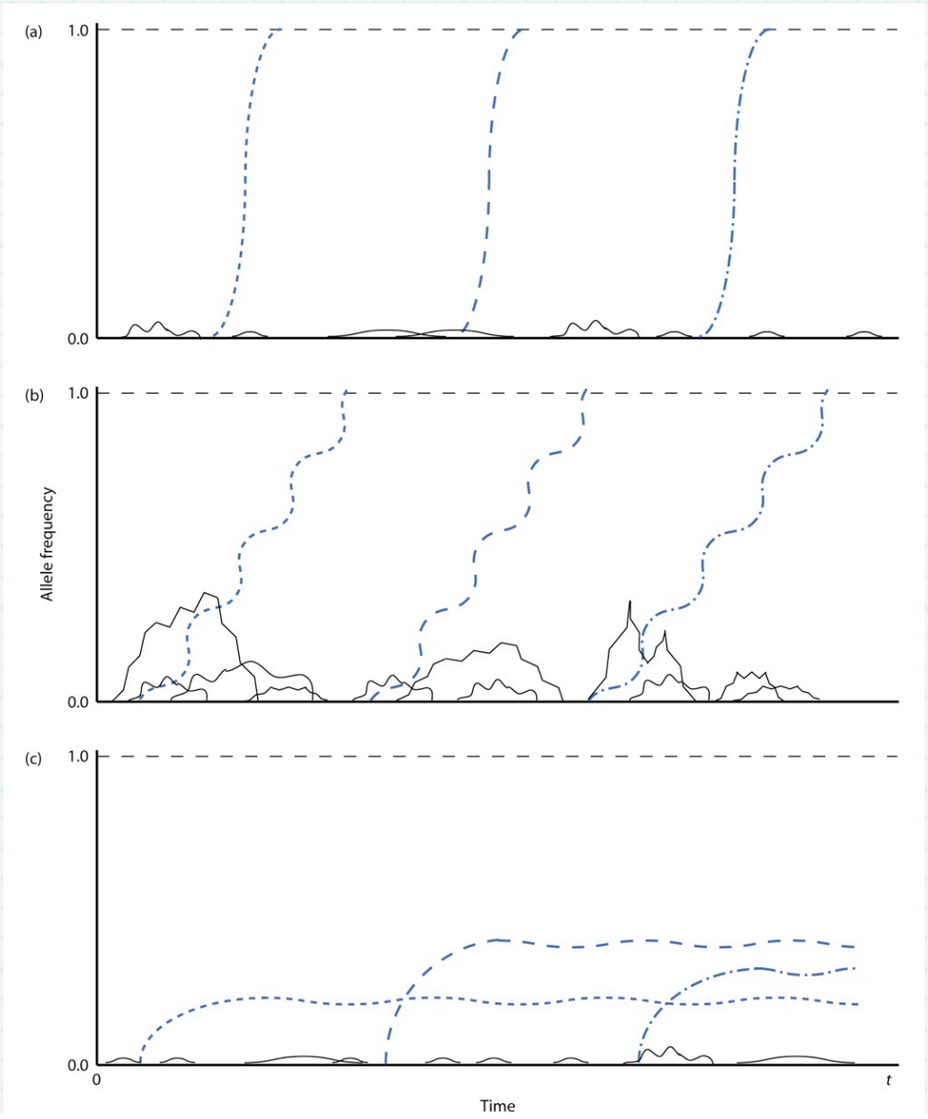
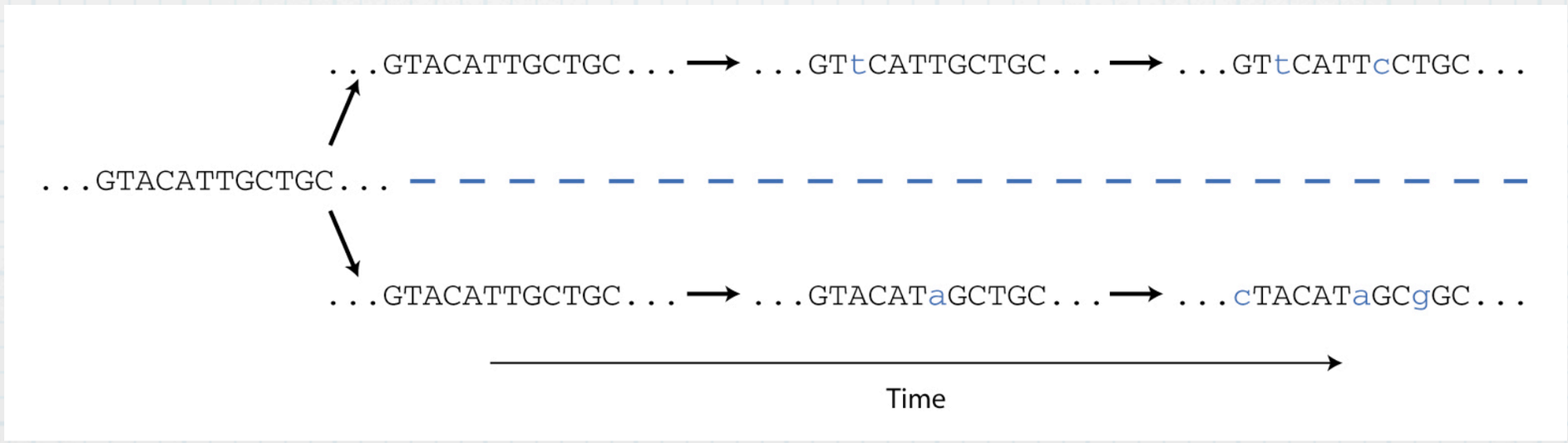


Figure 8.3 The dwell time for new mutations is different if fixation and loss is due to genetic drift or natural selection. With neutral mutations (b), most mutations go to loss fairly rapidly and a few mutations eventually go to fixation. For eventual fixation or loss of neutral mutations the path to that outcome is a random walk, implying that the time to fixation or loss has a high variance. For mutations that fix because they are advantageous (a), directional selection fixes them rapidly in the population. Therefore under directional selection alleles segregate for a shorter time and there is less polymorphism than with neutrality. For mutations that show overdominance for fitness, natural selection favoring heterozygote genotypes maintains several alleles in the population indefinitely. Therefore balancing selection greatly increases the segregation time of alleles and increases polymorphism compared to neutrality. Both cases of natural selection (a and c) are drawn to show negative selection acting against most new mutations. If new mutations are deleterious then the time to loss is very short and there is very little random walk in allele frequency since selection is nearly deterministic.

(1) The Neutral Theory

Divergence

* Genetic divergence occurs by **substitutions** that accumulate in two DNA sequences over time
* Substitution is the complete replacement of one allele previously most frequent in the population with another allele that originally arose by mutation



(1) The Neutral Theory

Divergence

* The neutral theory predicts the rate at which allelic substitutions occur and thereby the rate at which divergence occurs

* Predicting the substitution rate for neutral alleles requires knowing the probability that an allele becomes fixed in a population and the number of mutations that occur each generation

(1) The Neutral Theory

Divergence

* The rate at which alleles that originally entered the population as mutations go to fixation per generation is

$$k = (2N\mu)\frac{1}{2N}$$

which simplifies to

$$k = \mu$$

(1) The Neutral Theory

Divergence

* Notice that the rate of substitution is simply equal to the mutation rate and does not depend on Ne
* This makes sense because in smaller populations we have higher chance of fixation but fewer mutations while in larger population we have lower chance of fixation but more mutations
* In other words, the rate of input of new mutations and the chance of fixation due to genetic drift exactly balance out

(1) The Neutral Theory

Divergence

* The neutral theory also predicts that the substitutions that ultimately cause divergence should occur at a regular average rate
  * [For waiting time processes, the time between events is the reciprocal of the rate of events]
  * Since the rate of neutral substitution is μ, the expected time between neutral substitutions is $1/\mu$ generations

(1) The Neutral Theory

The nearly neutral theory

* The **nearly neutral theory** considers the fate of new mutations if some portion of new mutations are acted on by natural selection of different strengths

* The nearly neutral theory recognizes three categories of new mutations: **neutral mutations**, **mutations acted on strongly by either positive or negative natural selection**, and **mutations acted on weakly by natural selection relative to the strength of genetic drift**

(1) The Neutral Theory

The nearly neutral theory

* For a new mutation in a finite population that experiences natural selection, the forces of directional selection and genetic drift oppose each other
* We have seen that (i) $1/(2N_e)$ quantifies the "push" on a new mutation towards fixation caused by genetic drift, and (ii) $2s$ (s is the selection coefficient) quantifies the force of selection on a new mutation towards fixation

(1) The Neutral Theory

The nearly neutral theory

* Therefore, the conditions where genetic drift and natural selection have approximately equal influence on the fate of allele frequencies are given by

$$2s = \frac{1}{2N_e}$$

* When 2s is within an order of magnitude of $1/(2N_e)$, an allele can be described as net neutral or nearly neutral since natural selection and genetic drift are approximately equal forces dictating the probability of fixation of an allele

(1) The Neutral Theory

The nearly neutral theory

* We had seen before "Kimura's rule of thumb":
  * If $4N_e s << 1$, then selection is weak relative to sampling, and genetic drift will dictate allele frequencies
  * If $4N_e s >> 1$, then selection is strong relative to sampling, and natural selection will dictate allele frequencies
  * If $4N_e s \approx 1$, then allele frequencies are unpredictable

(1) The Neutral Theory

The nearly neutral theory

* More formally, assuming p is the allele frequency, N_e is the effective population size, and s is the selection coefficient assuming codominance, Kimura showed that the probability of fixation for a new mutation in a finite population is

$$P_{fixation} = \frac{1 - e^{-4N_e sp}}{1 - e^{-4N_e s}}$$

(1) The Neutral Theory

The nearly neutral theory

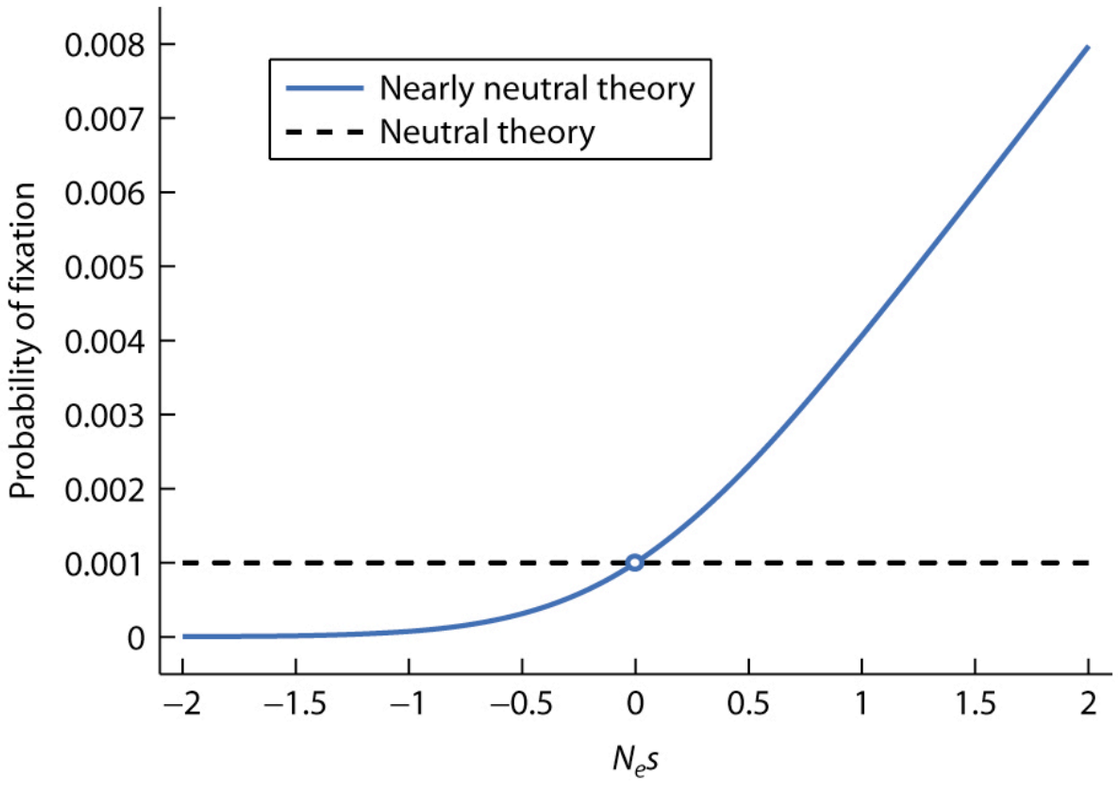


Figure 8.5 The probability of eventual fixation for a new mutation under the neutral and nearly neutral theories. Under the nearly neutral theory the probability of fixation depends on the balance between natural selection and genetic drift, expressed in the product of the effective population size and the selection coefficient ($N_e s$). When negative selection operates against a deleterious allele, the selection coefficient and $N_e s$ are negative. Values of $N_e s$ near 0 yield a fixation probability close to that predicted by neutral theory. Only when the absolute value of $N_e s$ is large does natural selection exclusively determine the probability of fixation. Neutral theory assumes that neutral mutations are not influenced by selection and have a constant probability of fixation dictated by the effective population size. In this example the initial allele frequency is 0.001, or the frequency of a new mutation at a diploid locus in a population of 500. After Ohta (1992).

(1) The Neutral Theory

The nearly neutral theory

* The nearly neutral theory predicts that the rate of substitution will depend on the effective population size for the proportion of mutations in a population that are nearly neutral ($4N_es \approx 1$)

* A consequence is that subdivided populations and different species can exhibit different levels of polymorphism based on their effective population size

* Similarly, rates of divergence can also vary between species, in contrast to the neutral theory, which predicts that the rate of substitution is independent of the effective population size

(2) Measures of Divergence and Polymorphism

* We will now introduce commonly used measures of divergence and polymorphism estimated from DNA sequence data

(2) Measures of Divergence and Polymorphism

DNA divergence between species

* To quantify molecular evolution by comparing two DNA sequences, the two sequences must first be **aligned** (more on this later), and then the number of sites that have different nucleotides is determined

* The number of nucleotide sites that differ between two sequences divided by the total number of sites compared gives the proportion of sites that differ, often called the **p distance**

* This is a basic measure of the evolutionary events that have occurred since the descent of the two sequences from a common ancestor

(2) Measures of Divergence and Polymorphism

DNA divergence between species

* The p distance between two DNA sequences sampled from completely independent populations should increase over time as substitutions within each population replace the nucleotide that was originally shared at each site due to identity by descent

* If the two DNA sequences represent two distinct species or completely isolated populations, then the p distance is a measure of divergence between the two species

(2) Measures of Divergence and Polymorphism

DNA sequence divergence and saturation

* **Saturation** is the phenomenon where DNA sequence divergence appears to slow and eventually reaches a plateau even as time since divergence continues to increase
* Saturation is caused by substitution occurring multiple times at the same nucleotide site, a phenomenon called **multiple hit** substitution
* In this case, the p distance is an under-estimate of the degree of divergence

(2) Measures of Divergence and Polymorphism

DNA sequence divergence and saturation

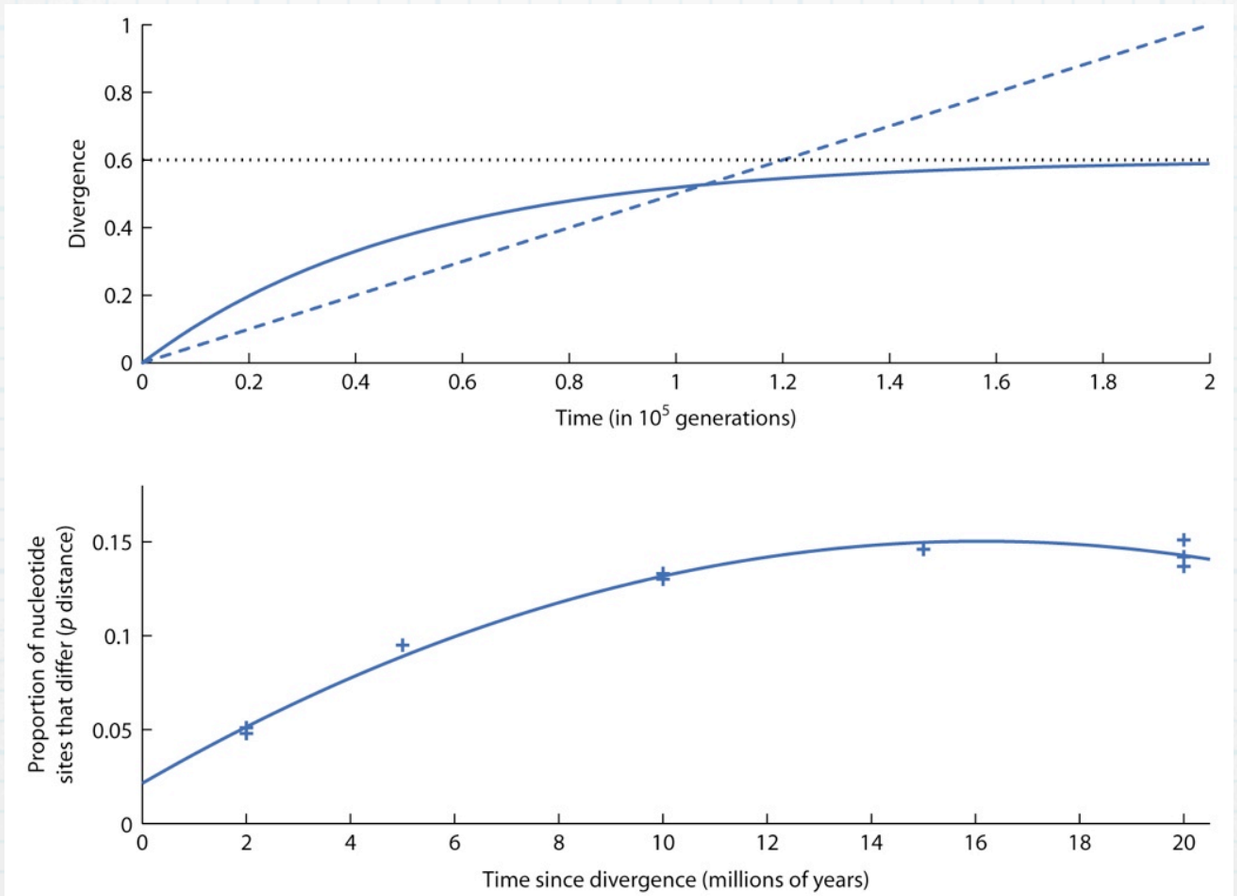


Figure 8.7 Substitutions that occur repeatedly at the same nucleotide site lead to saturation of nucleotide changes as time since divergence from a common ancestor increases. The rate of substitutions does not change and the total number of substitutions continues to increase over time, as shown by the dashed line in the top panel representing the true number of substitutions. In contrast, multiple substitutions at the same sites leads to a slowing and leveling off in the estimate of divergence (solid line, top panel). Therefore, the amount of divergence leads to the perception that the rate of divergence decreases over time. The bottom panel shows divergence and saturation at the mitochondrial cytochrome *c* oxidase subunit II gene among bovine species (ungulates including domestic cattle, bison, water buffalo, and yak) that diverged between 2 and 20 million years ago. In the top panel $\alpha = 1 \times 10^{-6}$ (α is explained overleaf). The bottom panel data are from Janecek et al. (1996) and the line is a quadratic regression fit.

(2) Measures of Divergence and Polymorphism

DNA sequence divergence and saturation

* At saturation, the two sequences can be viewed as random with respect to each other

* There are a wide variety of methods to **correct** the **perceived divergence** between two DNA sequences to obtain a better estimate of the **true divergence** after accounting for multiple hits

* These correction methods are called **nucleotide substitution models** and use parameters for DNA base frequencies and substitution rates to obtain a modified estimate of divergence

(2) Measures of Divergence and Polymorphism

DNA sequence divergence and saturation

* One of the simplest nucleotide substitution models is the **Jukes-Cantor** model

* This model assumes that any nucleotide in a DNA sequence is equally likely to be substituted with any of the other three nucleotides

* If we denote by α the probability of a nucleotide substitution, then the probability of any substitution is 3α

(2) Measures of Divergence and Polymorphism

DNA sequence divergence and saturation

* If the nucleotide is originally a G at generation zero, the probability that it is also a G one generation later is

$$P_{G(t=1)} = 1 - 3\alpha$$

* The probability of no substitutions over two generations is

$$P_{G(t=2)} = (1 - 3\alpha)^2$$

(2) Measures of Divergence and Polymorphism

DNA sequence divergence and saturation

* Another case where G could be observed two generations later is when in the first generation there was a substitution, and then a substitution back to G. The probability of this is

$$P_{G(t=2)} = \alpha(1 - P_{G(t=1)})$$

* Therefore, the probability of having G at generations zero and two is:

$$P_{G(t=2)} = (1 - 3\alpha)P_{G(t=1)} + \alpha(1 - P_{G(t=1)})$$

(2) Measures of Divergence and Polymorphism

DNA sequence divergence and saturation

* More generally, we have

$$P_{G(t+1)} = (1 - 3\alpha)P_{G(t)} + \alpha(1 - P_{G(t)})$$

* From this we can obtain

$$\Delta P_{G(t)} = (1 - 3\alpha)P_{G(t)} + \alpha(1 - P_{G(t)}) - P_{G(t)}$$

which simplifies to

$$\Delta P_{G(t)} = \alpha - 4\alpha P_{G(t)}$$

(2) Measures of Divergence and Polymorphism

DNA sequence divergence and saturation

* If we treat time as continuous, we get

$$\frac{\mathrm{d}P_{G(t)}}{\mathrm{d}t} = \alpha - 4\alpha P_{G(t)}$$

* The solution to this differential equation is

$$P_{G(t)} = \frac{1}{4} + (P_{G(t=0)} - \frac{1}{4})e^{-4\alpha t}$$

* As t gets large, $P_{G(t)}$ approaches 1/4

(2) Measures of Divergence and Polymorphism

DNA sequence divergence and saturation

* If the nucleotide at a site is initially G, then

$$P_{G(t)} = \frac{1}{4} + \frac{3}{4}e^{-4\alpha t}$$

* If the nucleotide at a site is initially not G, then

$$P_{G(t)} = \frac{1}{4} - \frac{1}{4}e^{-4\alpha t}$$

(2) Measures of Divergence and Polymorphism

DNA sequence divergence and saturation

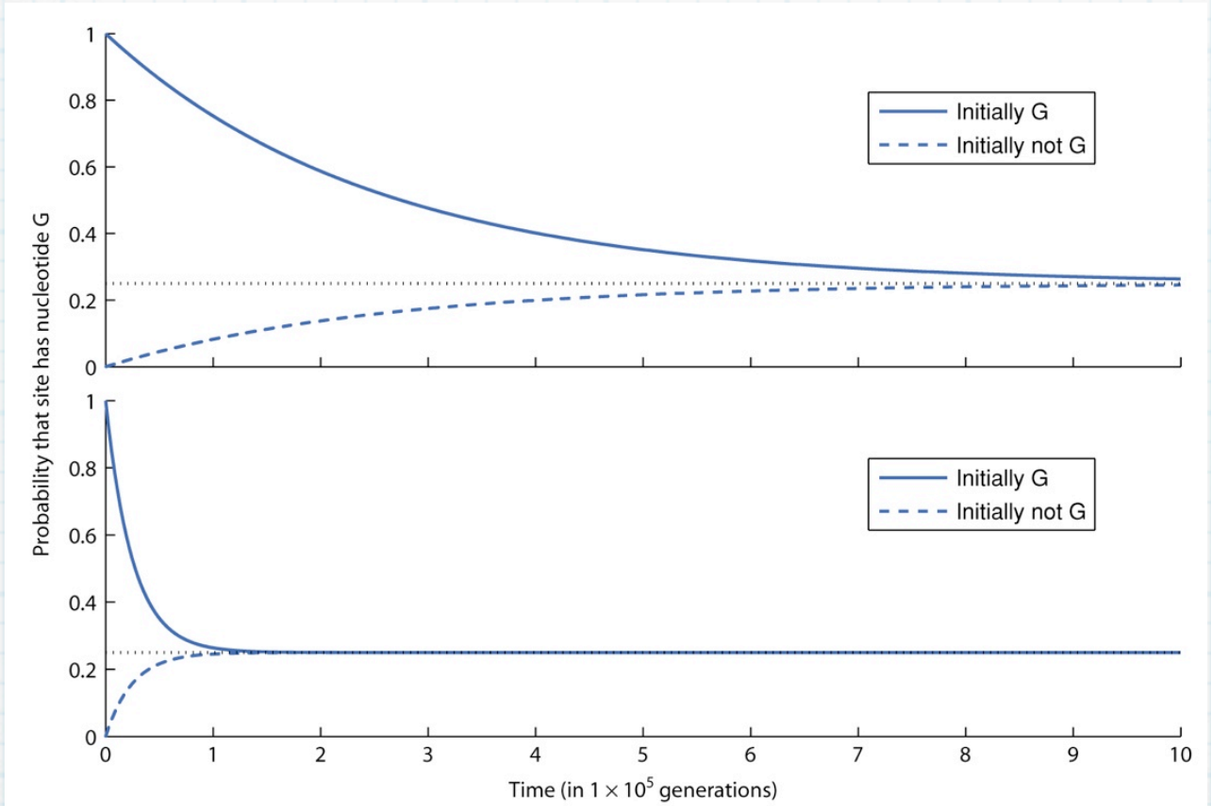


Figure 8.9 The probability that a nucleotide site retains its original base pair under the Jukes–Cantor model of nucleotide substitution. If a nucleotide site originally has a G base, for example, the probability of the same base being present declines steadily over time. If a nucleotide site was initially not a G (it was an A, C, or T), the probability that a G is present at the site increases over time. The probability that a given base is present always converges to 25% because that is the probability of sampling a given base at random if the probability of substitution to each nucleotide is equal. In the top panel $\alpha = 1 \times 10^{-6}$ whereas in the bottom panel $\alpha = 1 \times 10^{-5}$.

(2) Measures of Divergence and Polymorphism

DNA sequence divergence and saturation

* Given two DNA sequences that are identical by descent at time 0, at some later time t the probability that any site will possess the same nucleotide is

$$P_{I(t)} = \frac{1}{4} + \frac{3}{4}e^{-8\alpha t}$$

* The probability that two sites are different or divergent over time is $1-P_{I(t)}$, which is

$$d = \frac{3}{4}(1 - e^{-8\alpha t})$$

(2) Measures of Divergence and Polymorphism

DNA sequence divergence and saturation

* Therefore, we have

$$8\alpha t = -\ln\left(1 - \frac{4d}{3}\right)$$

* For two DNA sequences that were originally identical by descent, we expect that each site has a $3\alpha t$ chance of substitution

* Since there are two sequences, there is a $6\alpha t$ chance of a site being divergent between the two sequences

(2) Measures of Divergence and Polymorphism

DNA sequence divergence and saturation

* If we set expected divergence K=6αt, which is 3/4 of the expression 8αt, we get

$$K = -\frac{3}{4}\ln\left(1 - \frac{4d}{3}\right)$$

(2) Measures of Divergence and Polymorphism

DNA sequence divergence and saturation

* If we set expected divergence K=6αt, which is 3/4 of the expression 8αt, we get

$$K = -\frac{3}{4}\ln\left(1 - \frac{4d}{3}\right)$$

observed distance

(2) Measures of Divergence and Polymorphism

DNA sequence divergence and saturation

* If we set expected divergence K=6αt, which is 3/4 of the expression 8αt, we get

$$\boxed{K} = -\frac{3}{4}\ln\left(1 - \frac{4\boxed{d}}{3}\right)$$

corrected distance

observed distance

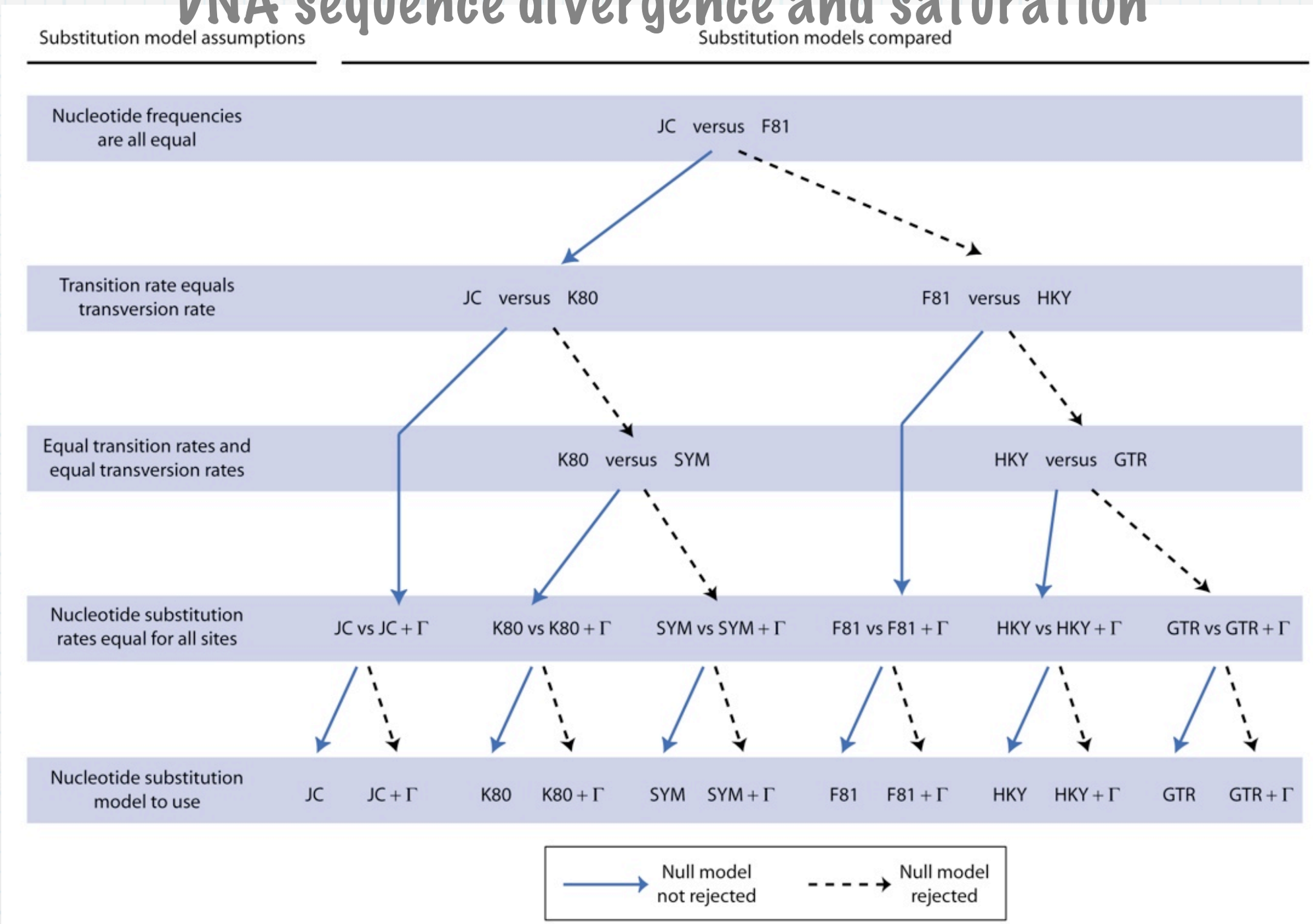
(2) Measures of Divergence and Polymorphism

DNA sequence divergence and saturation

* Jukes-Cantor is one of the simplest possible nucleotide substitution models
* Many other models are available to account for more complexities

(2) Measures of Divergence and Polymorphism

DNA sequence divergence and saturation



(2) Measures of Divergence and Polymorphism

DNA polymorphism

* One measure of DNA polymorphism is the **number of segregating sites**, S, in a set of aligned DNA sequences from one species
* A segregating site is one that exhibits at least two different nucleotides within the population
* The number of segregating sites per site is

$$p_S = \frac{S}{L}$$

(2) Measures of Divergence and Polymorphism

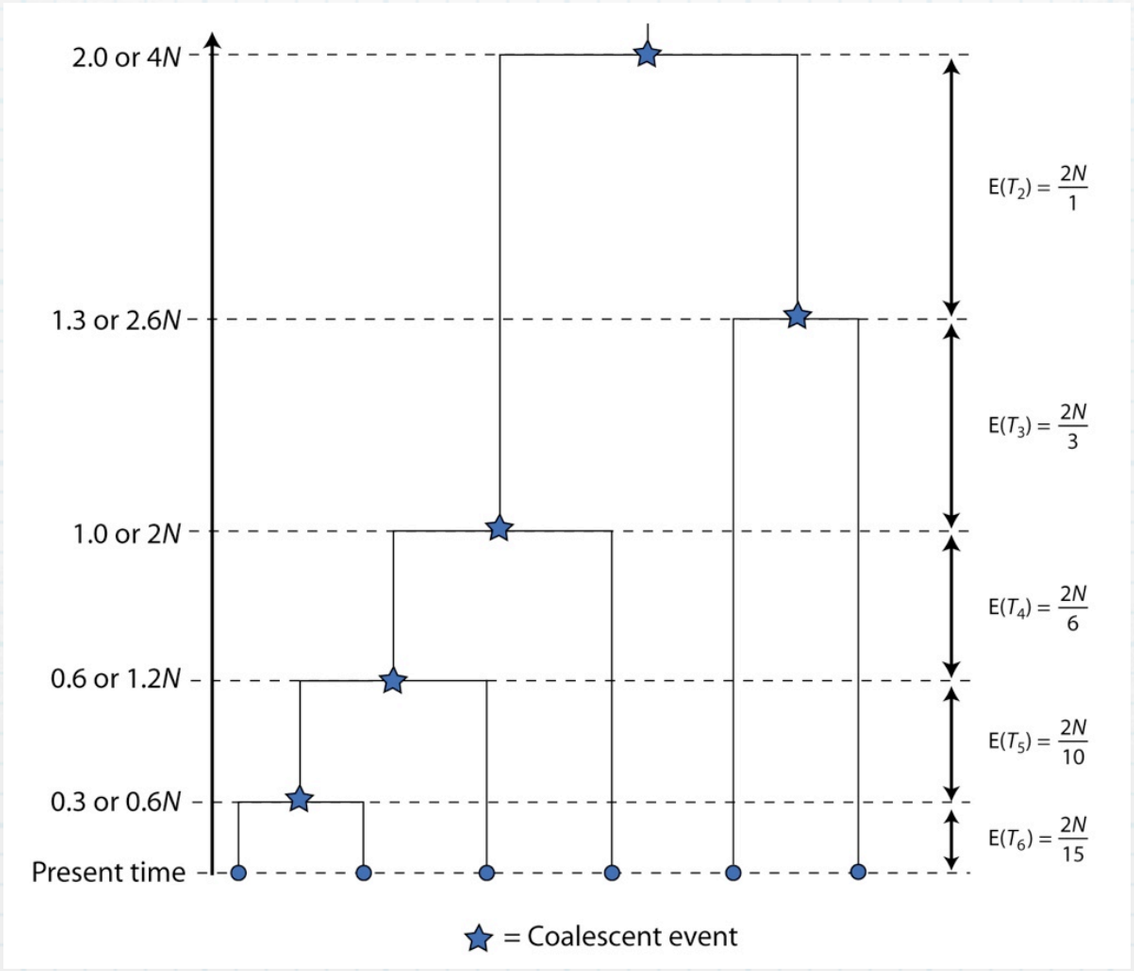
DNA polymorphism

* The number of segregating sites (S) under neutrality is a function of the scale mutation rate $\theta=4N_e\mu$
* One way to estimate θ under the infinite sites model is to consider the coalescent model

(2) Measures of Divergence and Polymorphism

DNA polymorphism

* Recall:



$$E(T_k) = \frac{2N_e}{\frac{k(k-1)}{2}} = \frac{2(2N_e)}{k(k-1)}$$

(2) Measures of Divergence and Polymorphism

DNA polymorphism

* Assuming n lineages in the present, the expected number of mutations is:

$$\mathbb{E}[S] = \mathbb{E}[\sum_{k=2}^{n} \mu k T_k] = \mu \sum_{k=2}^{n} k\mathbb{E}[T_k]$$

* Substituting $E(T_k)$ yields:

$$\mathbb{E}[S] = \mu \sum_{k=2}^{n} k \frac{2(2N_e)}{k(k-1)}$$

which simplifies to

$$\mathbb{E}[S] = 4N_e\mu \sum_{k=1}^{n-1} \frac{1}{k}$$

(2) Measures of Divergence and Polymorphism

DNA polymorphism

* Using $\theta=4N_e\mu$, we get:

$$\mathbb{E}[S] = \theta \sum_{k=1}^{n-1} \frac{1}{k}$$

* Rearranging gives:

$$\theta = \frac{\mathbb{E}[S]}{\sum_{k=1}^{n-1} \frac{1}{k}}$$

(2) Measures of Divergence and Polymorphism

DNA polymorphism

* Using $\theta=4N_e\mu$, we get:

$$\mathbb{E}[S] = \theta \sum_{k=1}^{n-1} \frac{1}{k}$$

* Rearranging gives:

$$\theta = \frac{\mathbb{E}[S]}{\sum_{k=1}^{n-1} \frac{1}{k}} \leftarrow a_1$$

(2) Measures of Divergence and Polymorphism

DNA polymorphism

$$\hat{\theta}_S = \frac{\mathbb{E}[S]}{a_1}$$

$$\hat{\theta}_S = \frac{p_S}{a_1}$$

(2) Measures of Divergence and Polymorphism

DNA polymorphism

* A second measure of DNA polymorphism is the **nucleotide diversity** in a sample of DNA sequences, also known as the **average pairwise differences**, denoted by π

* Given n sequences, and denoting by d_{ij} the number of sites that differ between sequences i and j, we define

$$\hat{\pi} = \frac{1}{\frac{n(n-1)}{2}} \sum_{i=1}^{n} \sum_{j>i}^{n} d_{ij}$$

(2) Measures of Divergence and Polymorphism

DNA polymorphism

* In larger samples that may include multiple identical DNA sequences, and denoting by p_i the frequency of sequence i, the nucleotide diversity can be estimated by

$$\hat{\pi} = \frac{k}{k-1} \sum_{i=1}^{k} \sum_{j>i}^{k} p_i p_j d_{ij}$$

(2) Measures of Divergence and Polymorphism

DNA polymorphism

* π is a measure of heterozygosity for DNA sequences

* As such, the value of π is a function of $\theta=4N_e\mu$ under an equilibrium between genetic drift and mutation

* With an estimate of π and the mutation rate at a locus (μ), it is possible to estimate the effective population size

(2) Measures of Divergence and Polymorphism

DNA polymorphism

Sequence 1	A	A	T	G	T	C	A	A	C G
Sequence 2	A	A	T	G	T	C	A	A	C G
Sequence 3	A	T	T	G	T	C	A	A	C G
Sequence 4	A	T	T	G	T	G	A	T	C G
		*				*		*	
Site number	1	2	3	4	5	6	7	8	9 10

Segregating sites (S and p_S):

Sites 2, 6, and 8 have variable base pairs among the four sequences (columns marked with *). These are segretating sites. Therefore, for these sequences $S = 3$ segregating sites and $p_S = 3/10 = 0.3$ segregating sites per nucleotide site examined.

Nucleotide diversity (π):

1 AATGTCAACG
2 AATGTCAACG $d_{12} = 0$

1 AATGTCAACG
3 ATTGTCAACG $d_{13} = 1$

2 AATGTCAACG
3 ATTGTCAACG $d_{23} = 1$

1 AATGTCAACG
4 ATTGTGATCG $d_{14} = 3$

2 AATGTCAACG
4 ATTGTGATCG $d_{24} = 3$

3 ATTGTCAACG
4 ATTGTGATCG $d_{34} = 2$

$$\Sigma d_{ij} = 0 + 1 + 3 + 1 + 3 + 2 = 10$$

Number of pairs of sequences compared = $[n(n-1)]/2 = [4(3)]/2 = 6$
$\hat{\pi}$ = 10 differences/6 pairs = 1.67 average pairwise differences
$\hat{\pi}$ = 1.67 avg. differences/10 sites = 0.167 pairwise differences per site

Figure 8.11 A hypothetical sample of four DNA sequences that are each 10 nucleotides long. There a total of three segregating sites ($S = 3$) or three-tenths of a segregating site per nucleotide ($p_S = 0.3$). The nucleotide diversity is calculated by summing the nucleotide sites that differ between each unique pair of DNA sequences. In this example there are 1.67 average pairwise nucleotide differences or 0.167 average pairwise nucleotide differences per nucleotide site.

(2) Measures of Divergence and Polymorphism

DNA polymorphism

Table 8.1 Nucleotide diversity (π) estimates reported from comparative studies of DNA sequence polymorphism from a variety of organisms and loci. All estimates are the average pairwise nucleotide differences per nucleotide site. For example, a value of $\pi = 0.02$ means that two in 100 sites vary between all pairs of DNA sequences in a sample.

Species	Locus	π	Reference
Drosophila melanogaster	*anon*1A3	0.0044	Andolfatto 2001
	Boss	0.0170	
	transformer	0.0051	
Drosophila simulans	*anon*1A3	0.0062	
	Boss	0.0510	
	transformer	0.0252	
Caenorhabditis elegans[a]	*tra-2*	0.0	Graustein et al. 2002
	glp-1	0.0009	
	COII	0.0102	
Caenorhabditis remanei[b]	*tra-2*	0.0112	
	glp-1	0.0188	
	COII	0.0228	
Arabidopsis thaliana[a]	*CAUL*	0.0042	Wright et al. 2003
	ETR1	0.0192	
	RbcL	0.0012	
Arabidopsis lyrata ssp. *Petraea*[b]	*CAUL*	0.0135	
	ETR1	0.0276	
	RbcL	0.0013	

[a]Mates by self-fertilization.
[b]Mates by outcrossing.

(3) DNA Sequence Divergence and the Molecular Clock

* One key result of the neutral theory is the prediction that the rate of substitution is equal to the mutation rate

* A corollary of this prediction is that the expected number of generations between substitutions is the reciprocal of the mutation rate

* Thus, the neutral theory provides a null model for the rate of divergence of genome regions between isolated populations called the **molecular clock hypothesis**

(3) DNA Sequence Divergence and the Molecular Clock

* The molecular clock hypothesis: the neutral theory prediction that divergence should occur at a constant rate over time so that the degree of molecular divergence between species is proportional to their time of separation

(3) DNA Sequence Divergence and the Molecular Clock

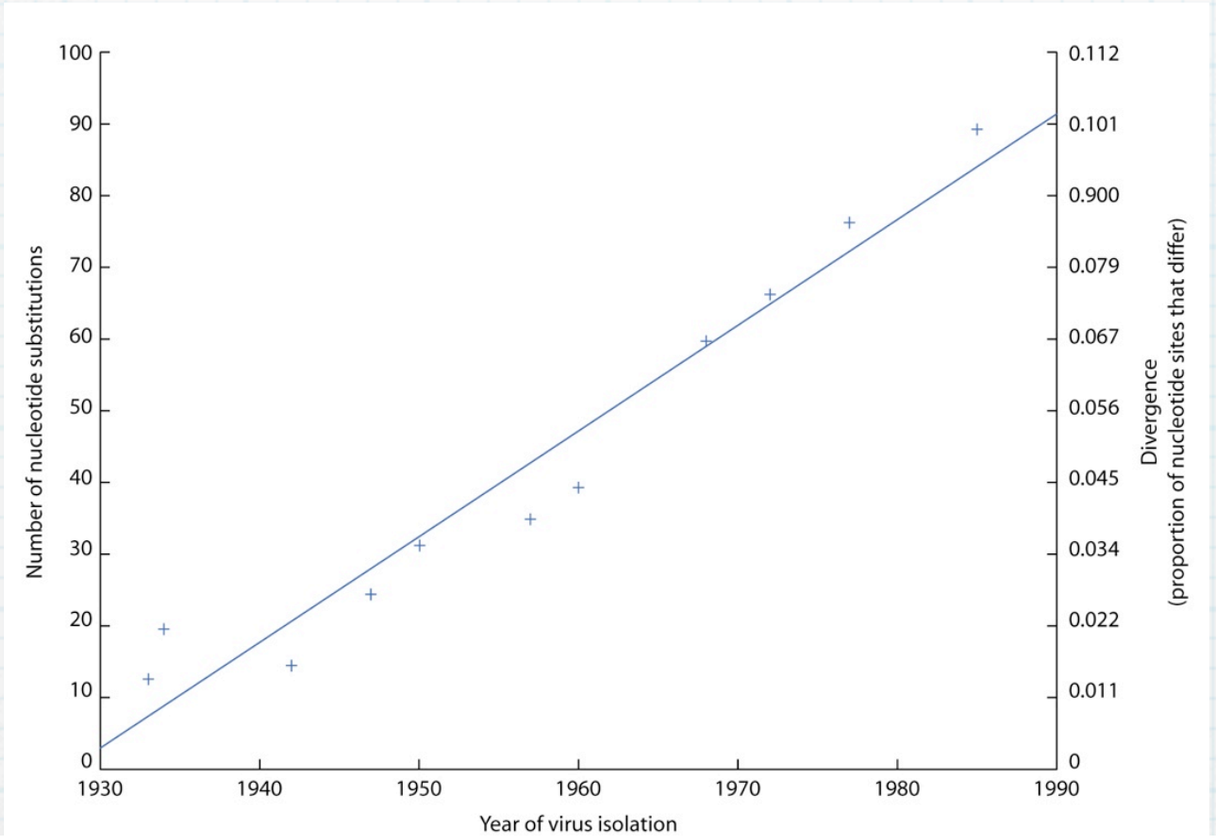


Figure 8.12 Rates of nucleotide change in the NS gene that codes for "nonstructural" proteins based on 11 human influenza A virus samples isolated between 1933 and 1985. The number of years since isolation and DNA sequence divergence from an inferred common ancestor are positively correlated. The pattern of increasing substitutions as time since divergence increases is expected under the molecular clock hypothesis. The observed rate of substitution was approximately 1.9×10^{-3} substitutions per nucleotide site per year, a very high rate compared to most genes in eukaryotes. The line is a least-squares fit. Data from Buonagurio et al. (1986).

(3) DNA Sequence Divergence and the Molecular Clock

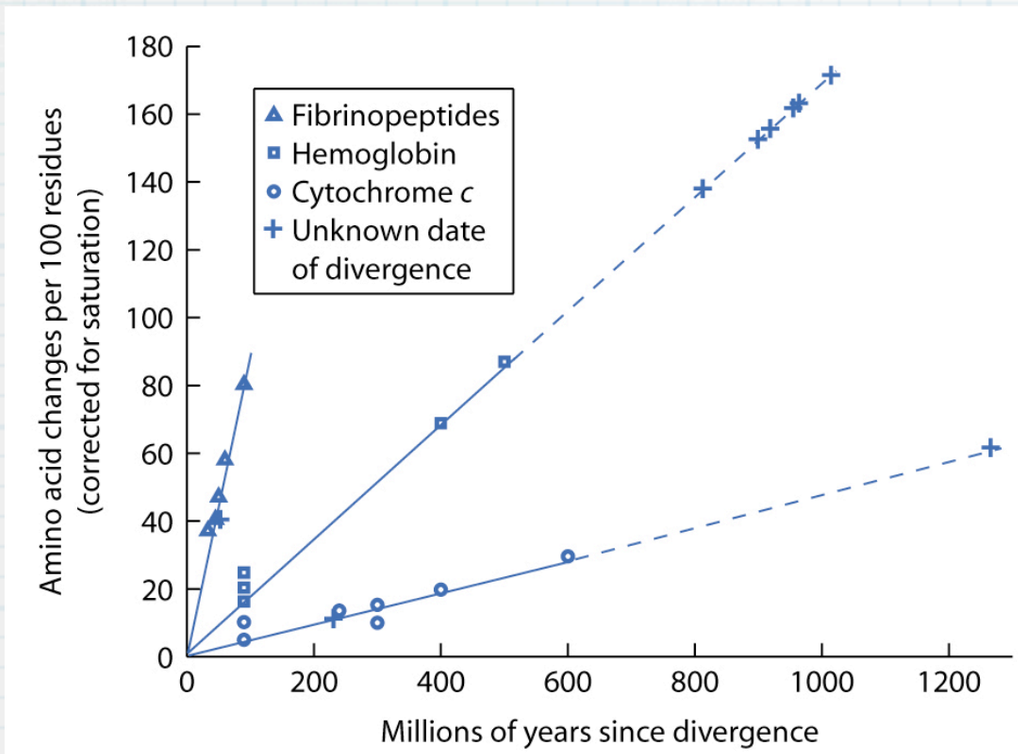


Figure 8.13 (*left*) Rates of protein evolution as amino acid changes per 100 residues in fibrinopeptides, hemoglobin, and cytochrome *c* over very long periods of time. Rates of divergence are linear over time for each protein, as expected for a molecular clock. Different proteins have different clock rates due to different mutation rates and degrees of functional constraint imposed by natural selection. Amino acid changes between pairs of taxa with unknown divergence times are plotted on dashed lines with the same slope as lines through points for taxa with estimated divergence times. The six points with unknown divergence times for hemoglobin represent divergence of ancestral globins into hemoglobins and myoglobins in the earliest animals, events that the molecular clock estimates to have happened between 1.1 billion and 800 million years ago. Data from Dickerson (1971).

(3) DNA Sequence Divergence and the Molecular Clock

* A useful application of the molecular clock is to date divergence events between species

$$T = \frac{k}{2\mu}$$

- T: the two species diverged T time units ago
- k: the rate of substitution
- μ: the mutation rate

(3) DNA Sequence Divergence and the Molecular Clock

* In the case of more than two sequences, knowledge of divergence times for certain pairs of species can help estimate divergence times for other pairs

(3) DNA Sequence Divergence and the Molecular Clock

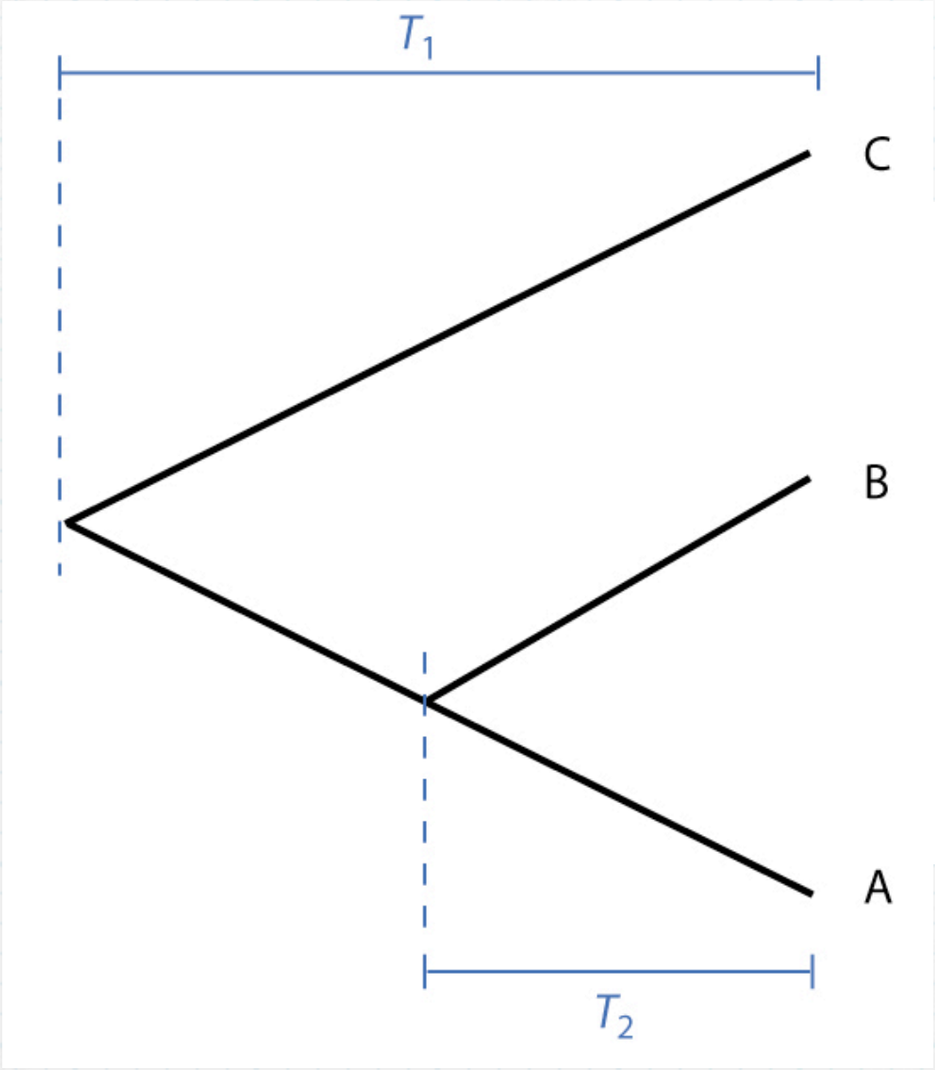


Figure 8.14 A schematic phylogenetic tree that can be used to date divergence events under the assumption of a constant rate of divergence over time or a molecular clock. T_1 is the time in the past when species C and the ancestor of species A and B diverged. T_2 is the time in the past when species A and B diverged. If either T_1 or T_2 are known, the rate of molecular evolution per unit of time can be estimated from observed sequence divergences. This rate of divergence can then be used to estimate the unknown amount of time that elapsed during other divergences.

Assuming T_1, K_{AB}, K_{AC}, and K_{BC} are all known, a we want to compute T_2

(3) DNA Sequence Divergence and the Molecular Clock

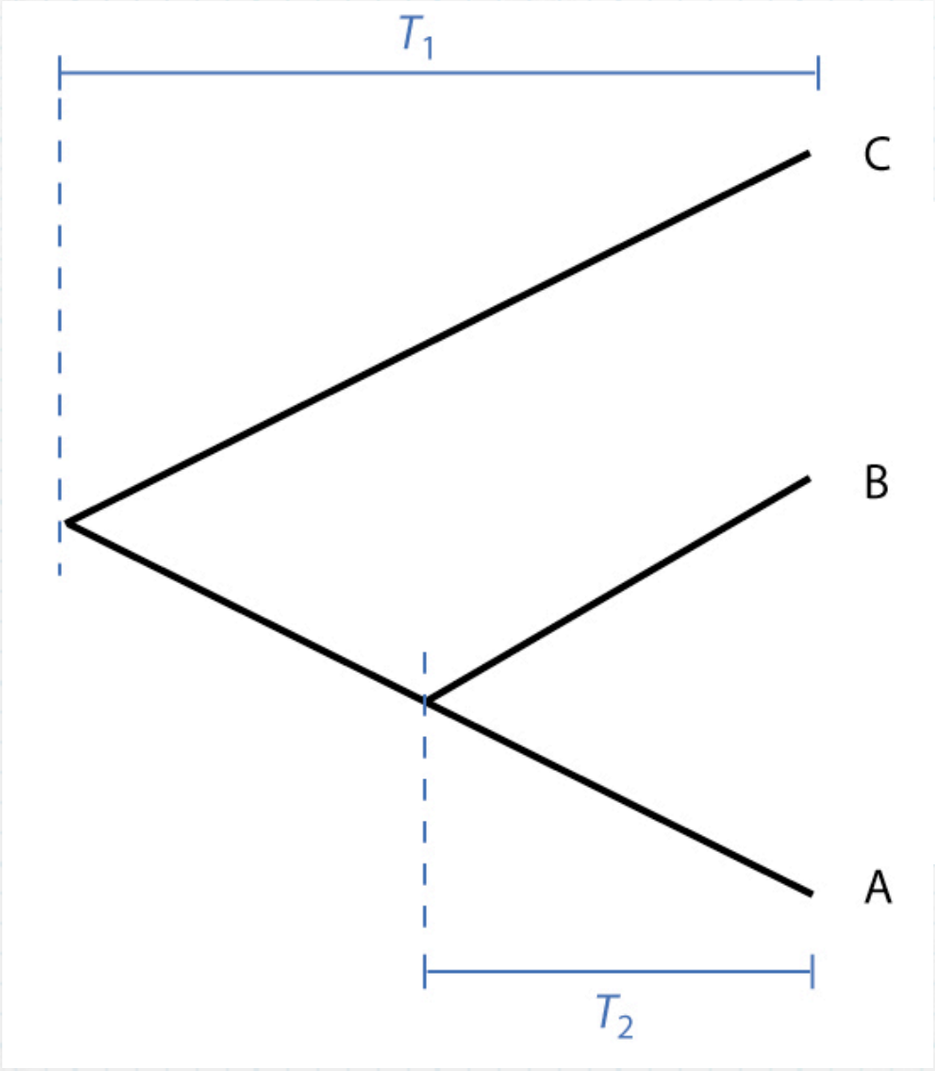


Figure 8.14 A schematic phylogenetic tree that can be used to date divergence events under the assumption of a constant rate of divergence over time or a molecular clock. T_1 is the time in the past when species C and the ancestor of species A and B diverged. T_2 is the time in the past when species A and B diverged. If either T_1 or T_2 are known, the rate of molecular evolution per unit of time can be estimated from observed sequence divergences. This rate of divergence can then be used to estimate the unknown amount of time that elapsed during other divergences.

Assuming T_1, K_{AB}, K_{AC}, and K_{BC} are all known, a we want to compute T_2

$$T_2 = \frac{K_{AB}}{2\mu}$$

(3) DNA Sequence Divergence and the Molecular Clock

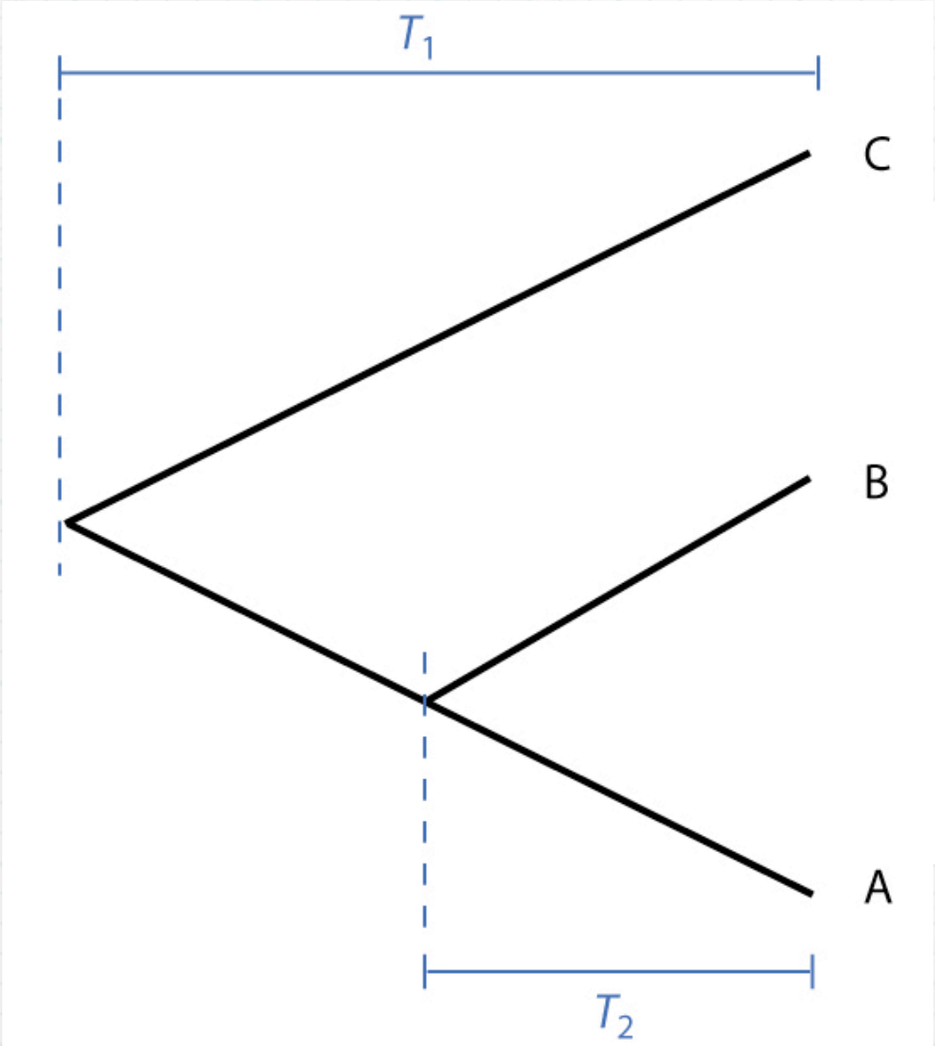


Figure 8.14 A schematic phylogenetic tree that can be used to date divergence events under the assumption of a constant rate of divergence over time or a molecular clock. T_1 is the time in the past when species C and the ancestor of species A and B diverged. T_2 is the time in the past when species A and B diverged. If either T_1 or T_2 are known, the rate of molecular evolution per unit of time can be estimated from observed sequence divergences. This rate of divergence can then be used to estimate the unknown amount of time that elapsed during other divergences.

Assuming T_1, K_{AB}, K_{AC}, and K_{BC} are all known, a we want to compute T_2

$$T_2 = \frac{K_{AB}}{2\mu} \quad + \quad \mu = \frac{1}{2}\left(\frac{K_{AC}}{2T_1} + \frac{K_{BC}}{2T_1}\right)$$

(3) DNA Sequence Divergence and the Molecular Clock

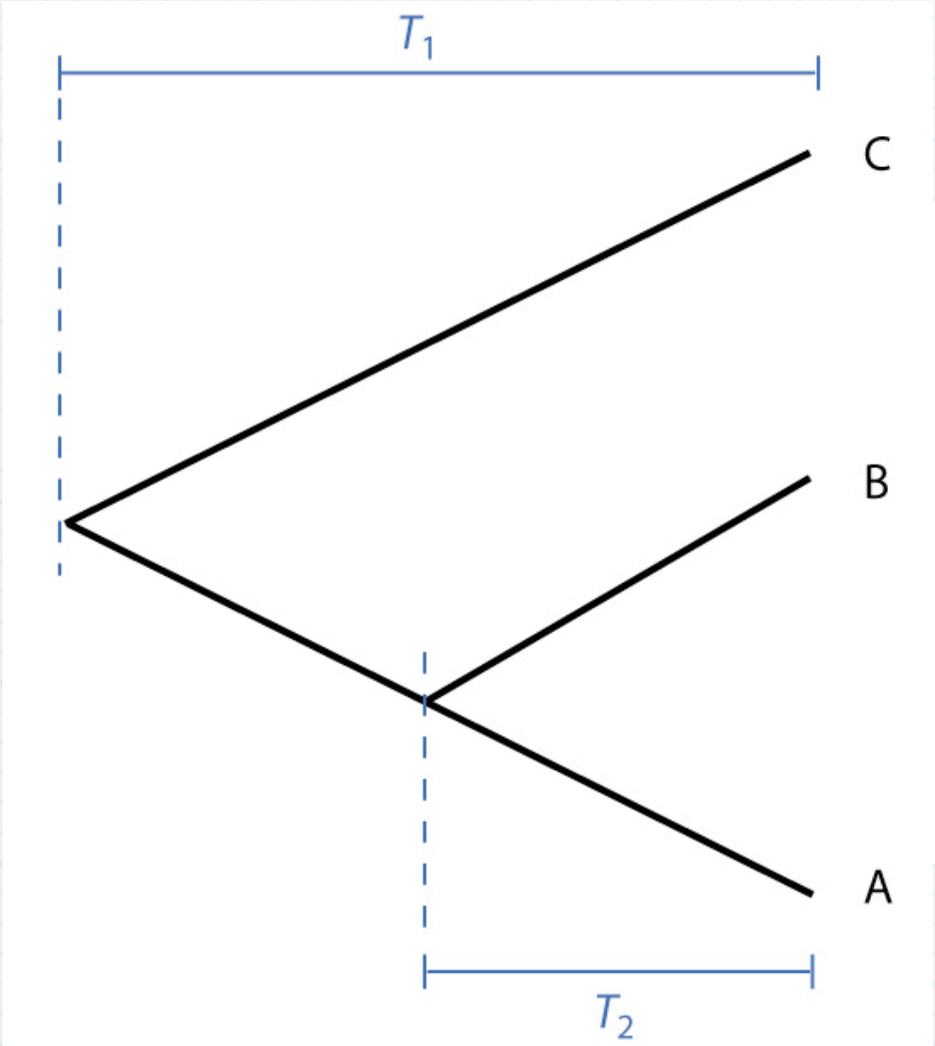


Figure 8.14 A schematic phylogenetic tree that can be used to date divergence events under the assumption of a constant rate of divergence over time or a molecular clock. T_1 is the time in the past when species C and the ancestor of species A and B diverged. T_2 is the time in the past when species A and B diverged. If either T_1 or T_2 are known, the rate of molecular evolution per unit of time can be estimated from observed sequence divergences. This rate of divergence can then be used to estimate the unknown amount of time that elapsed during other divergences.

Assuming T_1, K_{AB}, K_{AC}, and K_{BC} are all known, a we want to compute T_2

$$T_2 = \frac{K_{AB}}{2\mu} \quad + \quad \mu = \frac{1}{2}\left(\frac{K_{AC}}{2T_1} + \frac{K_{BC}}{2T_1}\right) \quad \Rightarrow \quad T_2 = \frac{2T_1 K_{AB}}{K_{AC} + K_{BC}}$$

(3) DNA Sequence Divergence and the Molecular Clock

* The molecular clock has been widely used to date major evolutionary transitions, establish times when the ancestors of many different organisms first evolved, and test hypotheses related to divergence times
* However, the use of the molecular clock to estimate times of divergence is complicated by numerous issues in practice

(3) DNA Sequence Divergence and the Molecular Clock

* Calibration times usually have considerable ranges, leading to uncertainty in any divergence time estimated from the molecular clock

* Corrections to divergence estimates are required for multiple substitutions occurring at the same nucleotide site

* The rate of substitution is assumed to be constant over time (variation in rates of substitution over time and across loci is now considered the rule rather than the exception!)

(4) Testing the Molecular Clock Hypothesis

* The molecular clock hypothesis provides a null model to examine the processes that operate during molecular evolution

* Rejecting the molecular clock hypothesis suggests that the sequences compared evolve at unequal rates (whether over time or among different lineages), a situation referred to as **rate heterogeneity**

(4) Testing the Molecular Clock Hypothesis

* Since the neutral theory leads to the molecular clock hypothesis, evidence for rate heterogeneity would appear to be evidence that genetic drift is not the main process leading to the ultimate substitution of most mutations

* Rejecting the hypothesis of rate homogeneity would suggest that natural selection is operating on mutations such that their rates of substitution are either sped up or slowed down relative to substitution rates under genetic drift

(4) Testing the Molecular Clock Hypothesis

* Originally, the molecular clock was proposed to model amino acid substitutions

* It was based on the Poisson process of substitutions, given by

$$P[N(t) \text{ substitutions at time t}] = \frac{e^{-\lambda t}(\lambda t)^{N(t)}}{N(t)!}$$

where λ is the rates of substitution per year

(4) Testing the Molecular Clock Hypothesis

* The Poisson model for a molecular clock implies that time intervals between substitutions are random in length

(4) Testing the Molecular Clock Hypothesis

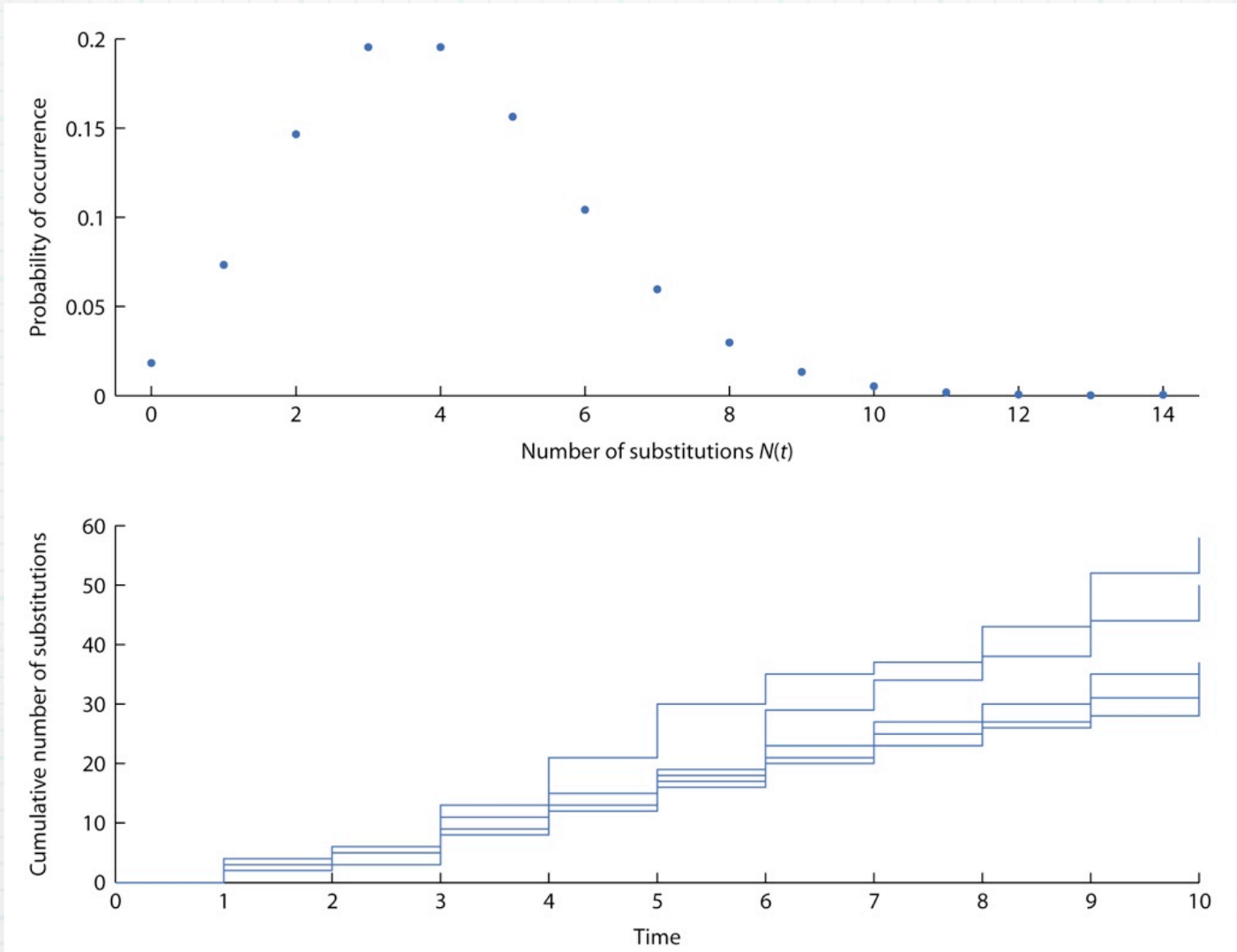


Figure 8.15 Substitution patterns under a Poisson process. The top panel shows the probability distribution for the number of substitutions that might occur during one time interval. $N(t)$ between 0 and 9 all have probabilities of greater than 0.01. The bottom panel shows the cumulative number of substitutions under a Poisson process for five independent trails. Each trail is akin to an independent lineage experiencing substitutions. The average number of substitutions is approximately 40 (four multiplied by the number of time intervals) but there is variation among the lineages. In both panels the rate of substitution is the same at $\lambda = 4$.

(4) Testing the Molecular Clock Hypothesis

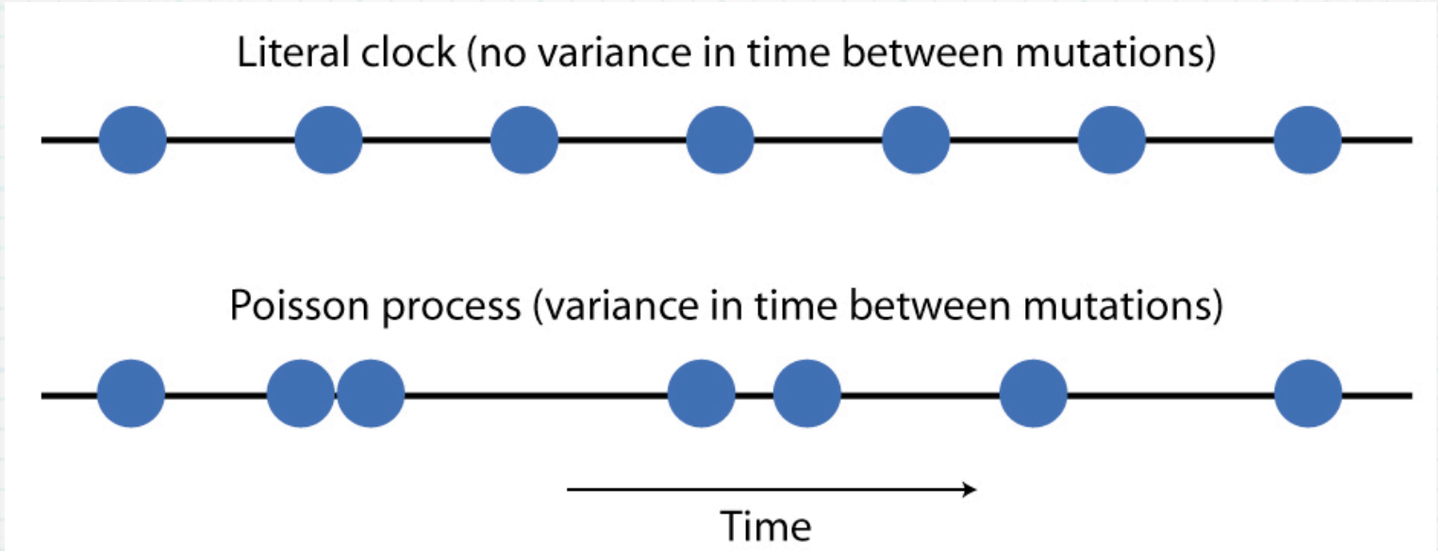
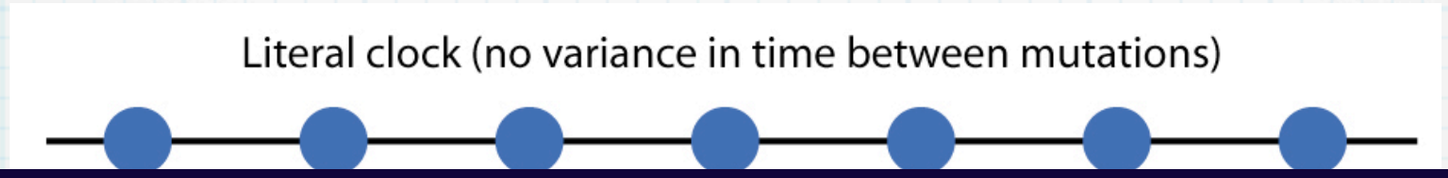


Figure 8.16 Two representations of rate at which substitution events (circles) occur over time. Mutations might occur with metronome-like regularity, showing little variation in the time that elapses between each mutation event. If substitution is a stochastic process, an alternative view is that the time that elapses between substitutions is a random variable. The Poisson distribution is a commonly used distribution to model the number of events that occur in a given time interval, so the bottom view is often called the Poisson molecular clock. Note that in both cases the number of substitutions and time elapsed is the same so that the average substitution rate is identical.

(4) Testing the Molecular Clock Hypothesis



A molecular clock that is based on a random process has inherent variation in the number of substitutions that occur over a given time interval even though the rate of substitution remains constant

Figure 8.16 Two representations of rate at which substitution events (circles) occur over time. Mutations might occur with metronome-like regularity, showing little variation in the time that elapses between each mutation event. If substitution is a stochastic process, an alternative view is that the time that elapses between substitutions is a random variable. The Poisson distribution is a commonly used distribution to model the number of events that occur in a given time interval, so the bottom view is often called the Poisson molecular clock. Note that in both cases the number of substitutions and time elapsed is the same so that the average substitution rate is identical.

(4) Testing the Molecular Clock Hypothesis

* Nonetheless, the Poisson process model of the molecular clock leads to a specific prediction about the variation in numbers of substitutions that should be observed if the rate of molecular evolution follows a Poisson process

* In particular, the Poisson distribution has the special property that the mean is equal to the variance

(4) Testing the Molecular Clock Hypothesis

* Therefore, the mean of and variance in the number of substitutions should be equal for independent DNA sequences evolving at the same rate according to a Poisson process

* The **index of dispersion** is used to compare the two quantities

$$R(t) = \frac{var(N(t))}{\mathbb{E}(N(t))}$$

(4) Testing the Molecular Clock Hypothesis

* If the numbers of substitutions in a sample of pairwise sequence divergences follow a Poisson process, we expect R(t)=1
* If R(t)>1, we have an **overdispersed molecular clock**

(4) Testing the Molecular Clock Hypothesis

* The molecular clock modeled as a Poisson process assumed it is possible to compare pairs of DNA sequences that were derived from a single DNA sequence in the past and then diverged instantly into two completely isolated species
* Actual DNA sequences usually have a more complex history that involves processes that operated in the ancestral species followed by the process of divergence in two separate species

(4) Testing the Molecular Clock Hypothesis

* In particular, in the ancestral species, the number and frequency of neutral alleles per locus were caused by population processes such as genetic drift and mutation

* This zone of **ancestral polymorphism** is the period of time when genetic variation in the ancestral species was dictated by drift-mutation equilibrium

(4) Testing the Molecular Clock Hypothesis

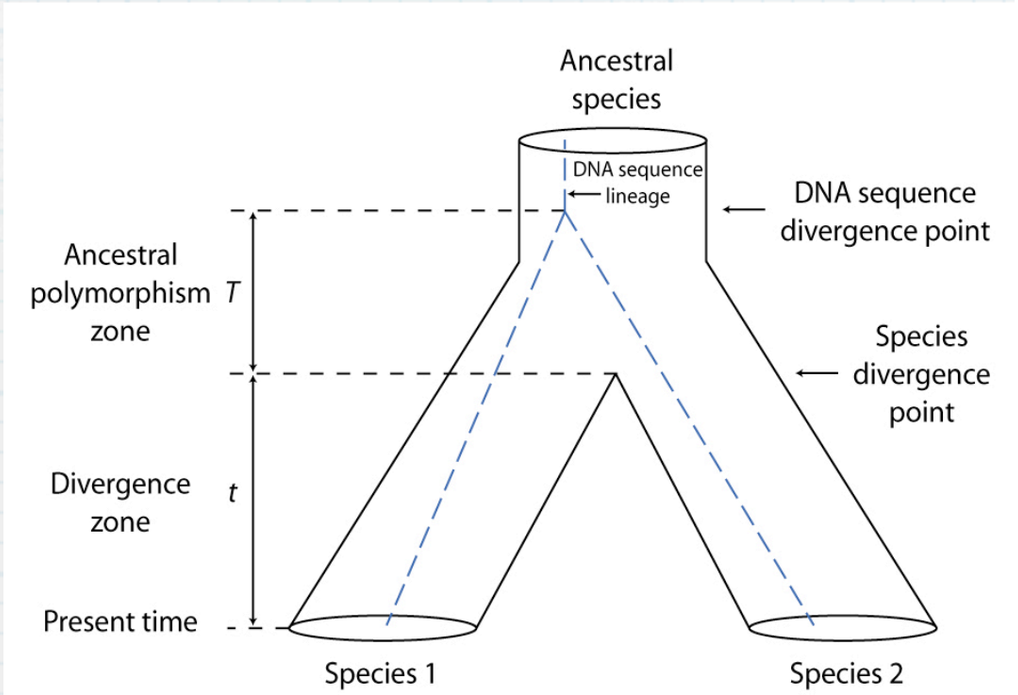


Figure 8.17 An illustration of the history of two DNA sequences that might be sampled from two species in the present time to estimate the rate of substitutions. The history is like a water pipe in an upside-down Y shape. The tube at the top contains the total population of lineages in the ancestral species, eventually splitting into populations of lineages that compose two species. The time when two lineages diverged from a common ancestor is not necessarily the same as the time of speciation. Therefore, a population process governing polymorphism operates for T generations in the ancestral species while a divergence process operates for t generations in the diverged species. The polymorphism process initially dictates the number of nucleotide changes between two sequences. Later, the divergence process dictates the number of nucleotide changes between two sequences. In two DNA sequences sampled in the present it is impossible to distinguish which process has caused the nucleotide changes observed.

(4) Testing the Molecular Clock Hypothesis

* The existence of both ancestral polymorphism and divergence processes complicates testing for overdispersion of the molecular clock
* Gillespie and Langley showed that a molecular clock combining polymorphism and divergence does not necessarily comprise a Poisson process where the index of dispersion is expected to equal one

(4) Testing the Molecular Clock Hypothesis

* Under these conditions, it was shown the dispersion index can be written as

$$R(t) = 1 + \frac{\theta^2}{\mathbb{E}(N(t))}$$

where $\theta = 4N_e\mu$

(4) Testing the Molecular Clock Hypothesis

* Ancestral polymorphism also presents difficulties for dating divergences using the molecular clock since sequence lineage history (genealogy) and species divergence history (species phylogeny) are not identical

* Estimates of time since divergence estimate the total elapsed time since the divergence of the two lineages rather than just the time since divergence of the two species

(4) Testing the Molecular Clock Hypothesis

* Ancestral polymorphism also presents difficulties for dating divergences using the molecular clock

The use of the molecular clock to date divergence time yields over-estimates of the species divergence time

not identical

* Estimates of time since divergence estimate the total elapsed time since the divergence of the two lineages rather than just the time since divergence of the two species

(5) Testing the Neutral Theory Null Model

* We'll describe four tests of the neutral theory null model:
  * The Hudson-Kreitman-Aguade (HKA) test
  * The McDonald-Kreitman (MK) test
  * Tajima's D statistic
  * Mismatch distributions

(5) Testing the Neutral Theory Null Model

The HKA test

* The HKA test compares neutral theory predictions for DNA sequence evolution with empirically estimated polymorphism and divergence

* The test requires DNA sequence data from two loci, one of which is chosen because it is selectively neutral and serves as a reference/control locus (e.g., non-coding regions or pseudo genes), while the other locus is the focus of the test and for which the neutral null model is being tested

(5) Testing the Neutral Theory Null Model

The HKA test

* Further, the HKA test requires that DNA sequence data for the two loci be collected in a particular manner:
  * DNA sequences for two loci must be obtained from two species to estimate divergence between the species for both loci
  * In addition, DNA sequences from multiple individuals within one of the species need to be obtained to estimate levels of polymorphism present at both loci

(5) Testing the Neutral Theory Null Model

The HKA test

* Polymorphism is measured by nucleotide diversity (π) for each locus

* Divergence is estimated by comparing the DNA sequences for both loci between an individual of each species, employing a nucleotide substitution model to correct for homoplasy

* Once these estimates are obtained, they can be combined

(5) Testing the Neutral Theory Null Model

The HKA test

Table 8.5 Estimates of polymorphism and divergence for two loci sampled from two species that form the basis of the HKA test. (a) The correlation of polymorphism and divergence under neutrality results in a constant ratio of divergence and polymorphism between loci independent of their mutation rate as well as a constant ratio of polymorphism or divergence between loci. (b) An illustration of ideal polymorphism and divergence estimates that would be consistent with the neutral null model. (c) Data for the *Adh* gene and flanking region (Hudson et al. 1987) is not consistent with the neutral model of sequence evolution because there is more *Adh* polymorphism within *Drosophila melanogaster* than expected relative to flanking region divergence between *D. melanogaster* and *D. sechellia*.

(a) Neutral case expectations

	Test locus	Neutral reference locus	Ratio (test/reference)
Focal species polymorphism (π)	$4N_e\mu_T$	$4N_e\mu_R$	$\frac{4N_e\mu_T}{4N_e\mu_R} = \frac{\mu_T}{\mu_R}$
Divergence between species (K)	$2T\mu_T$	$2T\mu_R$	$\frac{2T\mu_T}{2T\mu_R} = \frac{\mu_T}{\mu_R}$
Ratio (π/K)	$\frac{4N_e\mu_T}{2T\mu_T} = \frac{4N_e}{2T}$	$\frac{4N_e\mu_R}{2T\mu_R} = \frac{4N_e}{2T}$	

(b) Neutral case illustration

	Test locus	Neutral reference locus	Ratio (test/reference)
Focal species polymorphism (π)	0.10	0.25	0.40
Divergence between species (K)	0.05	0.125	0.40
Ratio (π/K)	2.0	2.0	

(c) Empirical data from *D. melanogaster* and *D. sechellia*

	Adh	*5' Adh* flanking region	Ratio (Adh/flank)
D. melanogaster polymorphism (π)	0.101	0.022	4.59
Between species divergence (K)	0.056	0.052	1.08
Ratio (π/K)	1.80	0.42	

(5) Testing the Neutral Theory Null Model

The HKA test

* The HKA test has some limitations and assumptions, which include:
  * the ability to identify an unambiguously neutral reference locus
  * the assumption that each of the two species used are panmictic

(5) Testing the Neutral Theory Null Model

The MK test

* The MK test is a test of the neutral model of DNA sequence divergence between two species
* The test requires DNA data from a single coding gene
* The sample of DNA sequences is taken from multiple individuals of a focal species to estimate polymorphism
* The test also requires a DNA sequence at the same locus from another species to estimate divergence

(5) Testing the Neutral Theory Null Model

The MK test

* The two classes of DNA changes used in the MK test are synonymous and nonsynonymous changes

(5) Testing the Neutral Theory Null Model

The MK test

Table 8.6 Estimates of polymorphism and divergence (fixed sites) for nonsynonymous and synonymous sites at a coding locus form the basis of the MK test. (a) Under neutrality, the number of nonsynonymous sites divided by the number of synonymous sites is equal to the ratio of the nonsynonymous and synonymous mutation rates. This ratio should be constant both for nucleotide sites with fixed differences between species and polymorphic sites within the species of interest. (b) An illustration of ideal nonsynonymous and synonymous site changes that would be consistent with the neutral null model. (c) Data for the *Adh* locus in *D. melanogaster* (McDonald & Kreitman 1991) show an excess of *Adh* nonsynonymous polymorphism compared with that expected based on divergence. (d) Data for the *Hla*-B locus for humans show an excess of polymorphism and more nonsynonymous than synonymous changes, consistent with balancing selection (Garrigan & Hedrick 2003).

	Fixed differences	Polymorphic sites
(a) Neutral case expectations		
Nonsynonymous sites (N)	$N_F = 2T\mu_N$	$N_P = 4N_e\mu_N$
Synonymous sites (S)	$S_F = 2T\mu_S$	$S_P = 4N_e\mu_S$
Ratio (N/S)	$\frac{N_F}{S_F} = \frac{2T\mu_N}{2T\mu_S} = \frac{\mu_N}{\mu_S}$	$\frac{N_P}{S_P} = \frac{4N_e\mu_N}{4N_e\mu_S} = \frac{\mu_N}{\mu_S}$
(b) Neutral case illustration		
Nonsynonymous changes	4	15
Synonymous changes	12	45
Ratio	0.33	0.33
(c) Empirical data from *Adh* locus for *D. melanogaster* (McDonald & Kreitman 1991)		
Nonsynonymous changes	~~2~~ 7	~~7~~ 2
Synonymous changes	~~42~~ 17	17 42
Ratio	~~0.045~~ 0.412	~~0.412~~ 0.048
(d) Empirical data for the *Hla*-B locus for humans (Garrigan & Hedrick 2003)		
Nonsynonymous changes	0	76
Synonymous changes	0	49
Ratio	–	1.61

(5) Testing the Neutral Theory Null Model

Tajima's D

* Tajima's D is a test of the standard coalescent model that is commonly applied to DNA polymorphism data sampled from a single species

* The test uses nucleotide diversity (π) and the number of segregating sites (S) observed in a sample of DNA sequences to make two estimates, $\hat{\theta}_\pi$ and $\hat{\theta}_S$, respectively, of the scaled mutation rate $\theta = 4N_e\mu$

(5) Testing the Neutral Theory Null Model

Tajima's D

* Tajima's D test relies on the fact that both estimates are expected to be approximately equal under the standard coalescent model where all mutations are selectively neutral and the population size remains constant over time

(5) Testing the Neutral Theory Null Model

Tajima's D

* The null hypothesis of the test is that the sample of DNA sequences was taken from a population with constant effective population size and selective neutrality of all mutations
* Natural selection operating on DNA sequences as well as changes in effective population size through time lead to rejection of this null hypothesis

(5) Testing the Neutral Theory Null Model

Tajima's D

* Tajima's D takes advantage of the fact that mutations that occurred further back in time in a genealogy are counted more times when computing the nucleotide diversity (π) than when computing the number of segregating sites (S)

(5) Testing the Neutral Theory Null Model

Tajima's D

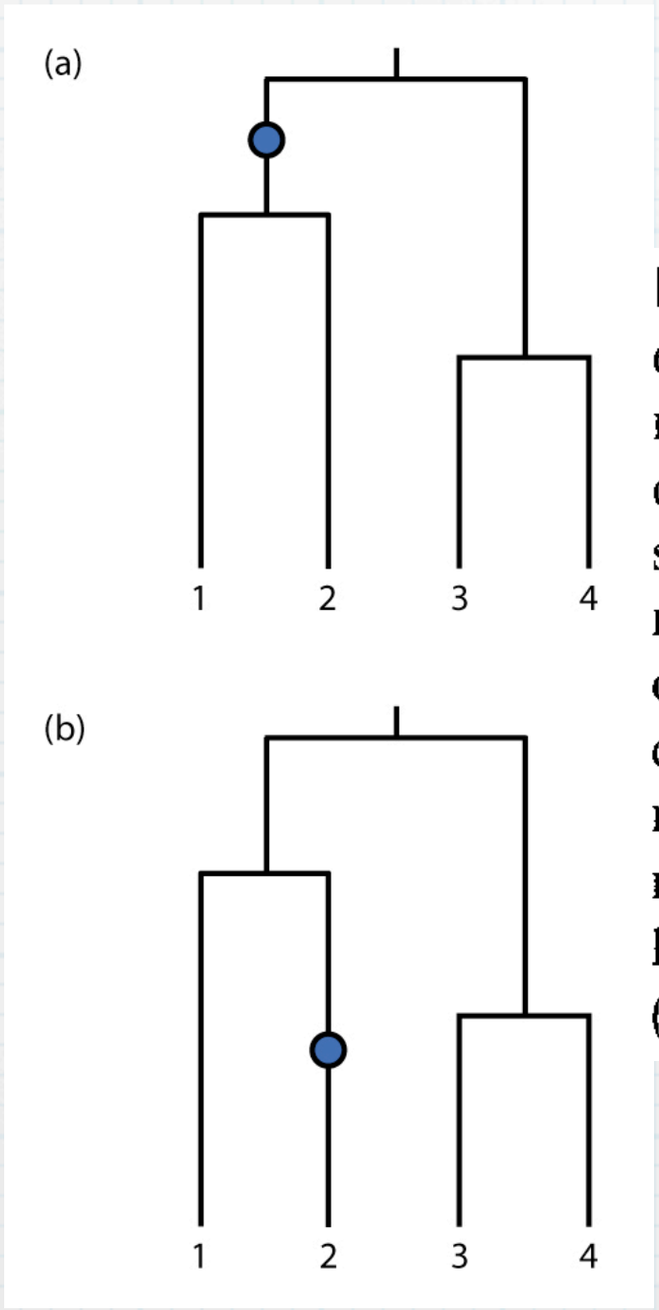


Figure 8.19 Estimates of the scaled mutation rate θ are estimated differently using nucleotide diversity ($\hat{\theta}_\pi$) and the number of segregating sites ($\hat{\theta}_S$) depending on the location of mutations in a genealogy. Each mutation makes a single segregating site under the assumptions of the infinite alleles model no matter where it occurs. However, mutations on internal branches will appear in multiple pairwise comparisons and cause π to be larger (a). In contrast, mutations that occur on external branches (b) that cause a nucleotide change in only a single lineage contribute less to π. Each mutation is counted four times (d_{13}, d_{23}, d_{14}, and d_{24}) in (a) but three times (d_{12}, d_{23}, and d_{24}) in (b) when computing π.

(5) Testing the Neutral Theory Null Model

Tajima's D

* The coalescent process with neutral alleles and constant N_e results in approximately the same total length along interior and exterior branches in a genealogy
* In contrast, processes that alter the probability of coalescence also change the ratio of interior and exterior branch lengths

(5) Testing the Neutral Theory Null Model

Tajima's D

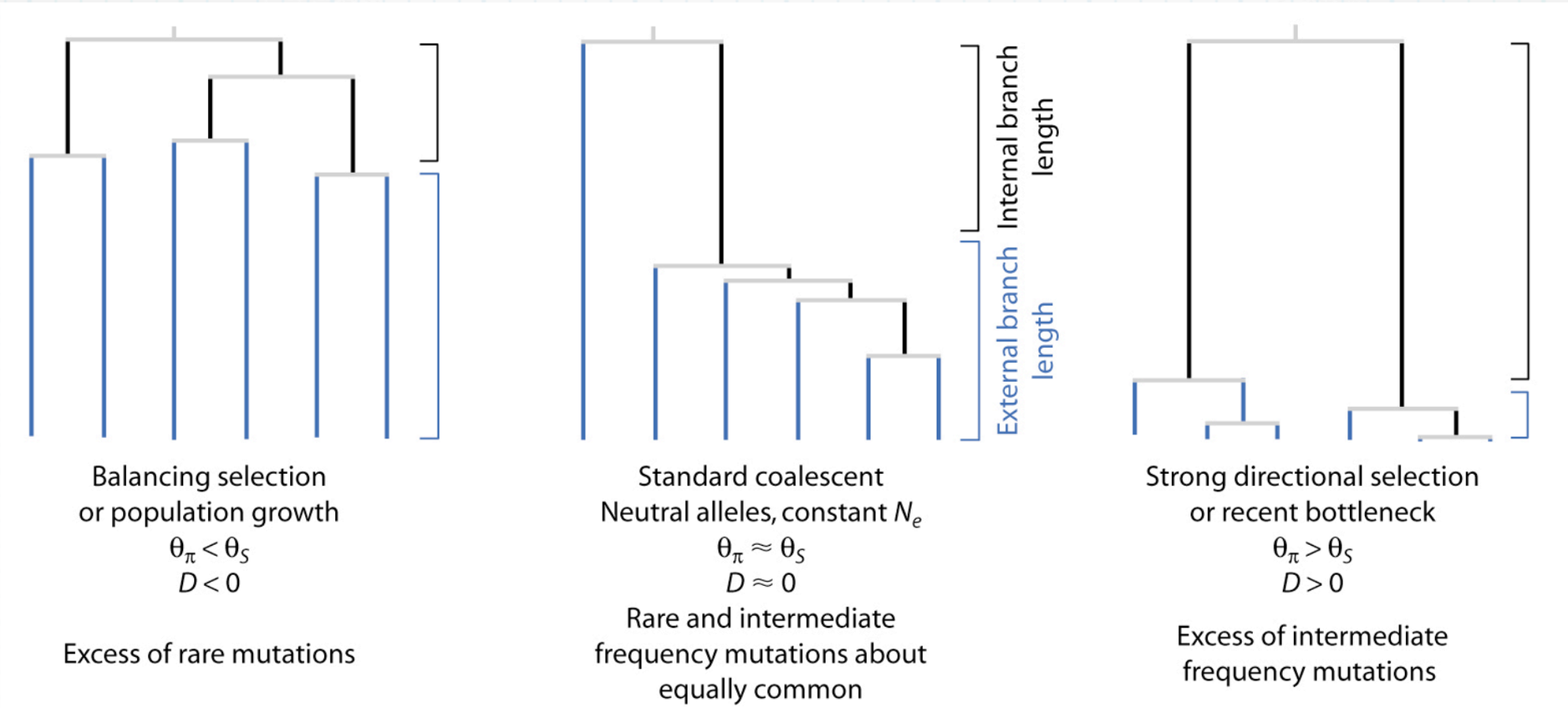


Figure 8.20 Differences in the shape of genealogies are the basis of Tajima's *D* test. In the standard coalescent model of genealogical branching the probability of coalescence is constant per lineage over time. The standard coalescent therefore gives expected branch lengths when all alleles are selectively neutral and the effective population size is constant (center). Changes in the effective population size over time (population growth, population bottlenecks) change the probability of coalescence over time as well. Natural selection also alters the probability of coalescence based on the fitness of alleles each lineage bears. Changes in the effective population size and natural selection alter the expected time to coalescence and therefore the expected branch lengths in a genealogical tree. If the chance of coalescence is greater in the present than in the past (right), most coalescent events occur near the present and internal branches are long in comparison with external branches. If the chance of coalescence is smaller in the present than in the past (left), most coalescent events occurred in the past and external branches are long in comparison with internal branches. Since the chance of a mutation is constant over time, lineages with longer branches are expected to experience more mutations.

(5) Testing the Neutral Theory Null Model

Tajima's D

* An alternative way to think about how Tajima's D works is to consider the frequency distribution of mutations under different types of natural selection or population histories

(5) Testing the Neutral Theory Null Model

Tajima's D

* Mutations that happen to occur on internal branches in a genealogy have an intermediate frequency

* Mutations that happen to occur on external branches in a genealogy have a low frequency

* Since the total internal and external branch length are expected to be about equal under the standard coalescent model, intermediate and rare alleles are also expected to be about equal in frequency

(5) Testing the Neutral Theory
Null Model
Tajima's D

* On the other hand,
  * both population growth and multi-allelic balancing selection can lead to an excess of rare mutations
  * strong purifying selection, shrinking population size, or a population bottleneck can lead to an excess of intermediate-frequency mutations

(5) Testing the Neutral Theory Null Model

Tajima's D

* These expectations are the basis for Tajima's D statistic

(5) Testing the Neutral Theory Null Model

Tajima's D

$$D = \frac{\hat{\theta}_\pi - \hat{\theta}_S}{\sqrt{var(\hat{\theta}_\pi - \hat{\theta}_S)}} = \frac{\hat{\theta}_\pi - \frac{p_S}{a_1}}{\sqrt{e_1 p_S + e_2 p_S (p_S - 1)}}$$

where

$$a_1 = \sum_{k=1}^{n-1} \frac{1}{k} \qquad a_2 = \sum_{k=1}^{n-1} \frac{1}{k^2} \qquad c = \frac{2(n^2 + n + 3)}{9n(n-1)} - \frac{n+2}{a_1 n} + \frac{a_2}{a_1^2}$$

$$e_1 = \frac{n+1}{3a_1(n-1)} - \frac{1}{a_1^2} \qquad e_2 = \frac{c}{a_1^2 + a_2}$$

n is the number of sequences sampled

(5) Testing the Neutral Theory Null Model

Tajima's D

* The null model is based on a constant mutation rate through time, the infinite sites model of mutation, the Wright-Fisher model with non-overlapping generations, and a panmictic population at drift-mutation equilibrium

* Although a large value of D serves to reject the standard coalescent model for a given set of DNA polymorphism data, distinguishing among the various causes of the rejection cannot be achieved with the estimate of D at a single locus

(5) Testing the Neutral Theory Null Model

Mismatch distributions

* So far, averages of nucleotide diversity and number of segregating sites have been used

* Alternatively, one can use the mismatch distribution directly, which is the frequency distribution of number of nucleotide sites that differ between all unique pairs of DNA sequences in a sample

(5) Testing the Neutral Theory Null Model

Mismatch distributions

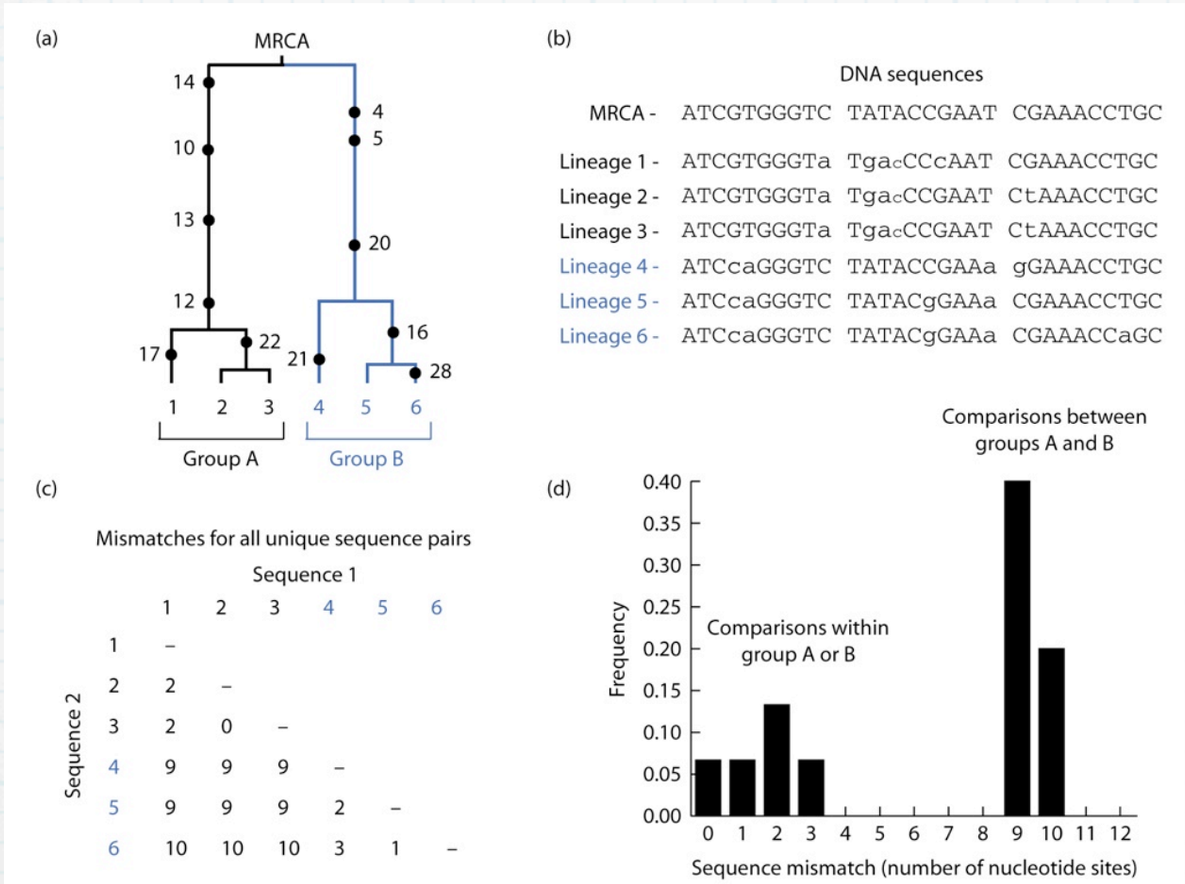


Figure 8.21 The basis of the mismatch distribution. (a) A neutral genealogy that bears multiple mutation events. Each mutation event is represented by a circle and the number of the random nucleotide site that mutated assuming the infinite sites mutation model. The six lineages in the present can be separated into two groups (called A and B) based on their ancestral lineage when there were only two lineages in the population. (b) The DNA sequences for each lineage are shown based on the 30 base-pair sequence assigned to the most recent common ancestor (MRCA) with mutations shown in lower-case letters. (c) The number of nucleotide sites that are different or mismatched between pairs of DNA sequences. (d) The mismatch distribution shown is a histogram of the mismatches for the 15 pairs of DNA sequences compared. Neutral genealogies from populations with constant N_e through time tend to show bimodal mismatch distributions. The cluster of observations with few mismatches results from sequence comparisons between recently related lineages (comparisons within group A or group B). In contrast, sequences from distantly related lineages that do not share the same ancestor when $k = 2$ (comparisons between groups A and B) tend to have more mismatches.

(5) Testing the Neutral Theory Null Model

Mismatch distributions

* A **bimodal mismatch distribution** is the characteristic signature of genealogies in populations with a relatively constant N_e in the past

* Significant deviations from this distribution may be indicative of processes that violate the underlying assumptions of the null hypothesis

(5) Testing the Neutral Theory Null Model

Mismatch distributions

* A related way to view polymorphism is by examining the **distribution of haplotype frequencies** in a sample of sequences

* Under neutrality and constant N_e, a range of haplotype frequencies are expected from very frequent to rare

* Processes that violate neutrality or change N_e over time may result in different haplotype frequency distributions

(5) Testing the Neutral Theory Null Model

Mismatch distributions

* Limitations of these tests include
    * Recombination obscures the history of mutations and in the extreme would lead to a uniform mismatch distribution
    * Given the stochastic nature of coalescence and the inherent large variance in times to coalescence, a large variance in the shape of the mismatch distribution may be observed even when N_e is constant

(6) Molecular Evolution of Loci That Are Not Independent

* When considering more than one locus, gametic disequilibrium influences polymorphism
* As natural selection drives a favorable mutation to high frequency in a population, the neutral alleles in gametic disequilibrium with the selected mutation also reach high frequency, a phenomenon that is called **genetic hitch-hiking**
* This also results in a loss of polymorphism in the population for the neutral alleles, since only those in gametic disequilibrium with the advantageous mutation remain in the population

(6) Molecular Evolution of Loci That Are Not Independent

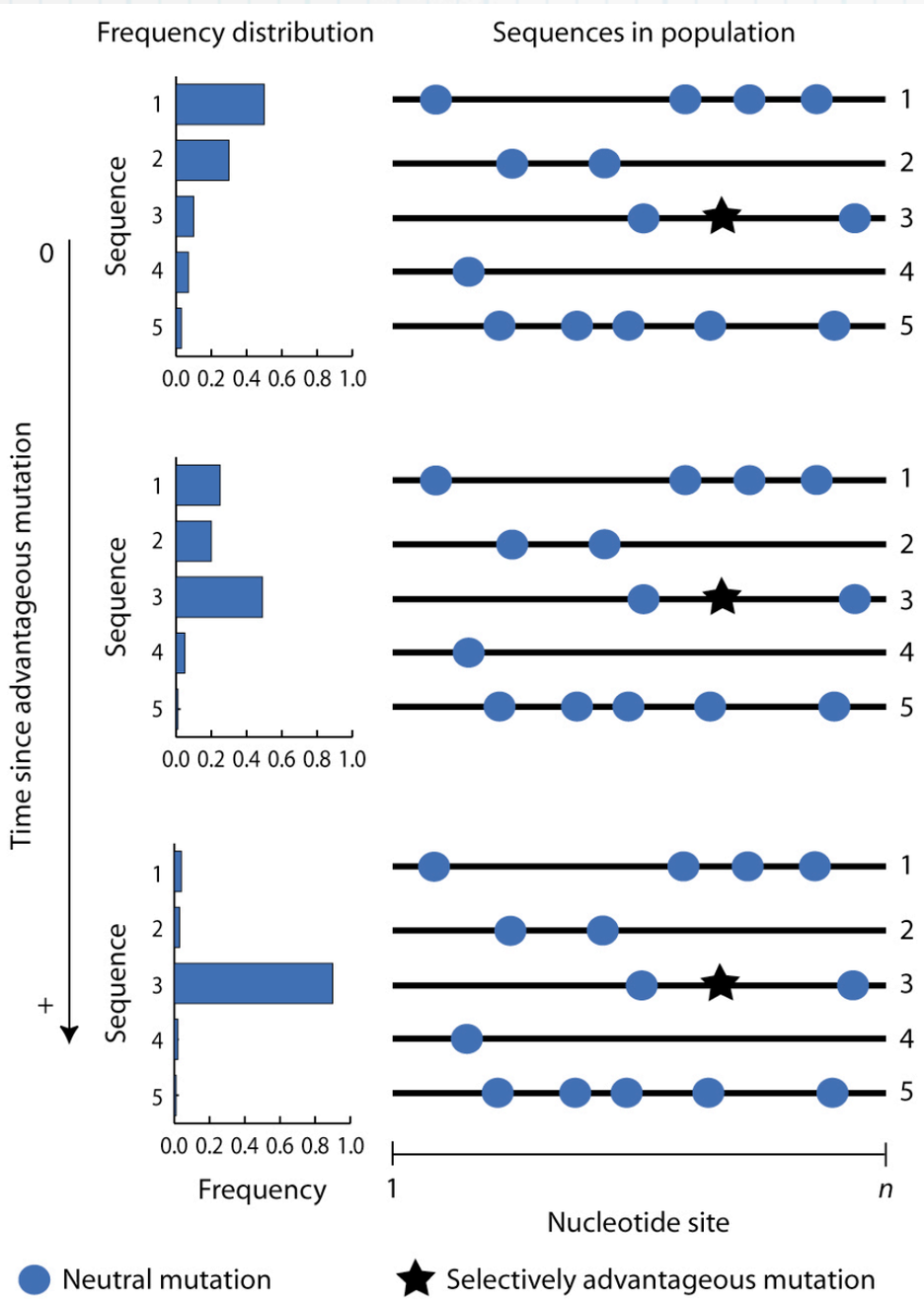


Figure 8.22 The impact of natural selection on an advantageous mutation as well as on associated nucleotide sites, often called a selective sweep. Imagine a single population that contains five distinct DNA sequences without recombination because reproduction is clonal. Each DNA sequence is distinguished by a number of neutral mutations (blue circles) and has a frequency given by the histogram on the left. Initially, the population has polymorphism since the population is composed of intermediate frequencies of each DNA sequence. At time 0, the third DNA sequence experiences a mutation that is strongly advantageous, indicated by the star. Natural selection acts to increase the frequency of the advantageous mutation over time, until the population approaches fixation for the third DNA sequence. Once selection has swept the advantageous mutation to near fixation the population has very little polymorphism. This is because only those original neutral mutations that were linked to the advantageous allele on the same DNA sequence remain in the population. Thus, positive selection on one site also sweeps away polymorphism at linked nucleotide sites if gametic disequilibrium is maintained. The figure assumes that positive natural selection is strong and increases the frequency of the third DNA sequence rapidly such that no new mutations appear in the population.

(6) Molecular Evolution of Loci That Are Not Independent

* The reduction in polymorphism caused by genetic hitch-hiking is sometimes called a **selective sweep**, since polymorphisms linked to an advantageous mutation are swept to high frequency by natural selections, and the other polymorphisms are swept out of the population at the same time

* Because mitochondrial genomes do not experience recombination, genetic hitch-hiking has the potential to cause strong selective sweeps

(6) Molecular Evolution of Loci That Are Not Independent

* It is also possible that new mutations may be deleterious
* In this case, it is expected that negative selection against deleterious mutations would reduce polymorphism through hitch-hiking in a process called **background selection**

(6) Molecular Evolution of Loci That Are Not Independent

* A third possibility is that new mutations are acted on by balancing selection, which would eventually bring new beneficial mutations at the same site to intermediate frequencies and maintain them in the population for very long periods of time

* Neutral sites in LD with sites under balancing selection have greatly increasing segregation times and so have a greater opportunity to experience mutation that leads to the accumulation of polymorphism

(6) Molecular Evolution of Loci That Are Not Independent

Gametic disequilibrium and rates of divergence

* We have seen that when the probability of fixation of a new mutation is dictated by natural selection, then divergence rates change for those sites under natural selection
* Positive natural selection speeds up divergence, and negative natural selection slows divergence rates
* Question: what happens to the divergence rates of neutral sites that are in LD with sites that are acted on by natural selection?

(6) Molecular Evolution of Loci That Are Not Independent

Gametic disequilibrium and rates of divergence

* We showed the rate of divergence is a function of the rate of substitution ($k=2T\mu$)

* The expected rate of substitution with a species is determined by the scaled mutation rate ($2N_e\mu=\theta/2$) and the probability of fixation for mutations (P_F), which can be stated as

$$k = \frac{\theta}{2} P_F$$

(6) Molecular Evolution of Loci That Are Not Independent

Gametic disequilibrium and rates of divergence

* For independent neutral mutations, we have

$$P_F = \frac{1}{2N}$$

* Hence, for such mutations, we have

$$k = \frac{\theta}{2} P_F = (2N_e\mu)\frac{1}{2N_e} = \mu$$

(6) Molecular Evolution of Loci That Are Not Independent

Gametic disequilibrium and rates of divergence

* Assume that we have a neutral locus with two alleles A and a, with frequencies x and 1-x, respectively

* Further, assume the neutral locus is completely linked to another locus where all alleles are under the influence of very strong positive natural selection

* Now, imagine a new mutation at the selected locus, which goes to fixation instantly

(6) Molecular Evolution of Loci That Are Not Independent

Gametic disequilibrium and rates of divergence

* What is the probability that the A allele at the neutral locus is also fixed due to hitch-hiking?

* The probability that the new advantageous mutation is linked to A is x

* Thus, there is the probability x that the A allele will sweep to fixation with the new mutation at the selected locus

(6) Molecular Evolution of Loci That Are Not Independent

Gametic disequilibrium and rates of divergence

* However, the probability that the A allele is fixed by genetic drift is also x, since that is the initial frequency of the A allele
* Therefore, even complete linkage to the selected allele does not alter the fixation probability for the A allele

Summary

* The neutral theory is a widely used null hypothesis in molecular evolution
* The neutral theory predicts that polymorphism is a function of N_e and the mutation rate
* The neutral theory predicts that the rate of divergence is a function of only the mutation rate
* Apparent divergence between two DNA sequences may be underestimated because of multiple hit mutations or homoplasy
* Nucleotide substitution models are used to correct observed divergence

Summary

* Nucleotide diversity (π) and the number of segregating sites (S) are two measures of DNA sequence polymorphism that can be used to estimate $\theta = 4N_e\mu$
* The molecular clock hypothesis uses the neutral theory prediction that divergence occurs at a constant rate over time to estimate the time that has elapsed since the two sequences shared a common ancestor
* The HKA and MK tests, Tajima's D, and the polymorphism frequency distribution can be used to test the null hypothesis
* Mutations acted on by natural selection can alter levels of polymorphism at linked neutral nucleotide sites